Using Importance Sampling to Simulate Queuing Networks with Heavy-Tailed Service Time Distributions

by

Karim Liman-Tinguiri

S.B. E.E.C.S. MIT 2010, S.B. Aerospace Eng. MIT 2010

Submitted to the Department of Electrical Engineering and Computer Science in Partial Fulfillment of the Requirements for the Degree of Master of Engineering in Electrical Engineering at the Massachusetts Institute of Technology

May 2011
[June 2011]

Author _____

Department of Electrical Engineering and Computer Science

May 11, 2011

Certified By _____

Eytan H. Modiano, Associate Professor

Thesis Supervisor

Accepted by _____

Dr. Christopher J. Terman

Chairman, Masters of Engineering Thesis Committee

Using Importance Sampling to Simulate Queuing Networks with Heavy-Tailed Service Time Distributions

by

Karim Liman-Tinguiri

Submitted to the Department of Electrical Engineering and Computer Science

May 10, 2011

In Partial Fulfillment of the Requirements for the Degree of Master of Engineering in Electrical Engineering

ABSTRACT

Characterization of steady-state queue length distributions using direct simulation is generally computationally prohibitive. We develop a fast simulation method by using an importance sampling approach based on a change of measure of the service time in an M/G/1 queue. In particular, we present an algorithm for dynamically finding the optimal distribution within the parametrized class of delayed hazard rate twisted distributions of the service time. We run it on a M/G/1 queue with heavy-tailed service time distributions and show simulation gains of two orders of magnitude over direct simulation for a fixed confidence interval.

Thesis Supervisor: Eytan H. Modiano
Title: Associate Professor

Contents

1 Introduction **6**

2 Direct Monte-Carlo and Importance Sampling **9**

2.1 Simulation framework 9

2.2 Direct Monte-Carlo simulation 12

2.3 Importance Sampling 14

2.4 Rare Events 18

2.5 Asymptotic optimality 25

3 Network Simulations **29**

3.1 Steady-state distributions 29

3.2 Importance sampling and GI/GI/1 queues 33

3.3 The M/G/1 queue 43

3.4 Delayed Hazard Rate Twisting 46

4 Adaptive Delayed Hazard Rate Twisting **51**

4.1 Single-parameter adaptive importance sampling 51

4.2 Adaptive plain hazard rate twisting 55

4.3 Multiple-parameter adaptive importance sampling 65

4.4 Adaptive delayed hazard rate twisting 67

5 Conclusion **71**

A Confidence Intervals **72**

B Heavy-tailedness **76**

C Optimal twisting of the M/M/1 queue 79

Acknowledgments

I would like to thank the many people without which this thesis would not have been possible. First, I would like to thank my dad Kiari, my mom Absatou, my brother Zacharie and my sister Aïda. They have all provided me with tremendous support and encouragements, and I certainly could not have done it without them.

I would also like to thank my research supervisor, Prof. Eytan H. Modiano, for his guidance during this project, and his patience with me as I shared with him myriads of far-fetched research ideas. His enthusiasm has kept me motivated throughout this project, and his thorough knowledge of the literature has saved me weeks of working on previously addressed problems. I could not have done it without him.

Lastly, I would like to thank my lab mates, Gregory Kuperman, Guner Celik, Marzieh Parandehgheibi, Matt Johnston, Nathan Jones, and Sebastian Neumayer for keeping me company and sharing useful advice with me as I transitioned from being an undergraduate to a graduate student.

1 Introduction

Obtaining analytical results for the performance of realistic network models proves to be challenging for even the simplest topologies. As a result, the only option available to compute estimates of various performance and robustness metrics such as buffer overflow probability is often simulation. Because the events of interest often occur with very minute probability, direct Monte Carlo simulation requires prohibitive computational resources and places challenging requirements on the random number generator.

The difficulties associated with direct Monte Carlo estimation of low probability events has led to the development of variance reduction techniques such as importance sampling. Importance Sampling is a family of estimation methods whereby the small probability event is made to happen much more often in simulation by using a different probability distribution. The over-estimated probability is then weighed by a likelihood ratio to yield an unbiased estimate of the probability of interest in the original system (see figure 1). When applied correctly, importance sampling can decrease by several orders of magnitude the number of samples required to estimate, within a specified confidence interval, the probability of an event compared to direct Monte Carlo simulation[1].

Good importance sampling methods exist for estimating buffer overflow probabilities in stable GI/GI/1 queues when the inter-arrival times and the service time distributions have tails that decrease exponentially or faster. Unfortunately, many real-life networks have been shown empirically to be

[1]See [10] for example, table II.

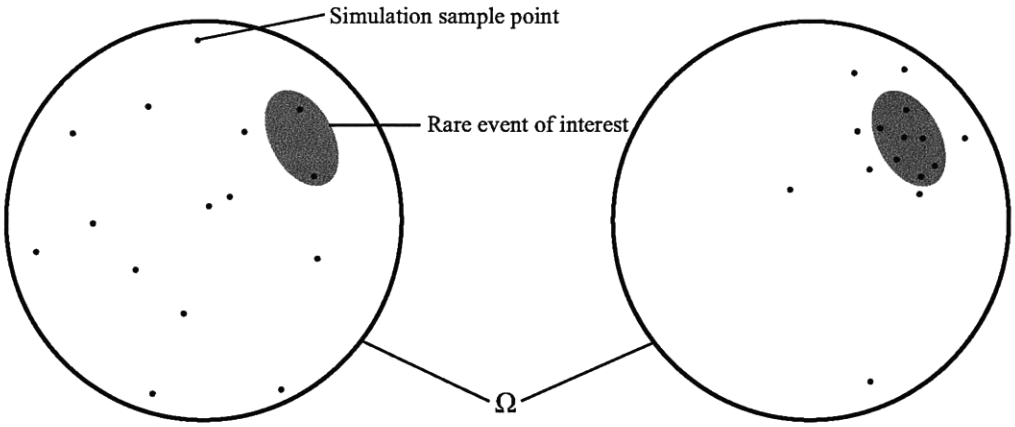

Figure 1: Side by side comparison of a direct Monte Carlo simulation (left) and an importance sampling one (right). In the direct Monte Carlo simulation, sample points are drawn according to the specified probability distribution, and the maximum likelihood estimate of the probability of the event ($\hat{p} = \frac{2}{15}$ in this example) is used. Contrast this with the importance sampling simulation where a different probability distribution is used to draw the sample points, one that makes the event of interest much more likely to occur. In order for the estimate of the probability to be unbiased, each occurrence of the event is weighed down by an appropriate likelihood ratio.

much better modeled with polynomially decreasing ("heavy") tails than with exponential ones [15]. One method, weighted delayed hazard rate twisting, has been shown to be efficient for simulating M/G/1 systems with heavy tails, but it requires solving a complicated equation that does not have a closed form for certain distributions.

Furthermore, provably efficient importance sampling estimators have been difficult to obtain because of analytical complications introduced when moving from a single queue in isolation to a network of queues. In this thesis, we propose a novel adaptive importance sampling algorithm based on hazard rate twisting where the twisting parameters are estimated dynamically during the simulation. We derive bounds for its performance and compare it with existing methods.

This thesis is organized as follows. In chapter 2, we examine the direct Monte Carlo simulation method and show that it is prohibitively expensive for low probability events. We then introduce the method of importance sampling and the performance criteria used to evaluate importance sampling estimators. In chapter 3, we focus on applying importance sampling to queuing systems. We present the key results from the literature, including the delayed hazard rate twisting estimator for heavy-tailed service time distributions. We touch upon difficulties faced in transitioning from queues to networks of queues and present some known results. Finally, in chapter 4, we present the novel adaptive hazard rate twisting algorithm together with some bounds on its performance. We run an implementation of it and compare its empirical performance with existing importance sampling algorithms.

2 Direct Monte-Carlo and Importance Sampling

When modeling systems with complex interactions, it is often the case that we do not have closed-form expressions for parameters of interest. For example in a network of stable interacting queues with known inter-arrival time and service time distribution, there is no general expression for the mean waiting time of a customer at any queue. One tool that can give us answers in these situations is simulation. The premise on which simulation algorithms are based is the law of large numbers, which can be loosely stated in this context as: "if we replicate a system *enough times* and average some output of that system, our empirical average is likely to be *very close* to that output's probabilistic mean."

This statement is purely qualitative and tells us little about the performance of our simulation so that we cannot readily compare one simulation method to another. The purpose of this chapter is to build an analytical framework rigorous enough to enable us to quantify the accuracy / computational cost trade-off of different simulation methods, and will culminate in the introduction of the method of importance sampling upon which this thesis's contribution is built.

2.1 Simulation framework

We consider a probability space with a sample space Ω, a probability measure P defined over some appropriate set of events, and a random vector $X : \Omega \to \mathbb{R}^n$. We are interested in using simulation to evaluate the mean of a real-valued random variable $h(X)$ for some specified function $h : \mathbb{R}^n \to \mathbb{R}$.

The vector X does not need to capture the entire state of the system, or more formally, it is not necessary for the random variable $X : \Omega \to \mathbb{R}^n$ to be an invertible mapping. Rather, X should be the most succinct representation possible of the system, subject to the following two constraints:

1. The metric we are interested in evaluating must be expressible as the mean of a random variable $h(X)$ for some specified function $h(\cdot)$, of the vector X, and

2. We must have a means to draw independent and identically distributed (i.i.d.) samples distributed according to the same distribution as X on a digital computer. A sufficient condition for this to be met is that we know explicitly the CDF of X. In that case, we may use the inversion method to draw i.i.d. samples with that distribution[2].

Definition 2.1. *A simulation problem is a quadruplet* (Ω, X, F_X, h) *where* Ω *is a probabilistic sample space,* $X : \Omega \to \mathbb{R}^n$ *is a random vector defined on* Ω, F_X *is the cumulative distribution function associated with* X *and* $h : \mathbb{R}^n \to \mathbb{R}$ *is a function that satisfies* $p = E[h(X)]$ *for some real number* p.

This framework is general enough to allow a wide range of metrics in stochastic systems to be expressed by choosing an appropriate function $h(\cdot)$ in the context of a system characterized by its sample space, and the distribution of X. As an example, suppose that our system consists of 100 independent dice rolls, each with an outcome independently and identically

[2]See [6] for more information on the inversion method in the case where X is a random variable. The extension to a random vector is trivial and omitted for brevity.

distributed (albeit not necessarily uniformly) between 1 and 6 so that the sample space can be defined as:

$$\Omega = \{(x_1, x_2, ..., x_{100}) \,|\forall i \in \{1, 2, ..., 100\}, x_i \in \{1, 2, 3, 4, 5, 6\}\}$$

i.e., the set of vectors of length 100 comprised of integer scalars between 1 and 6. Furthermore, suppose that we are interested in evaluating the probability p that die number 2 has a value greater than or equal to that of die number 1. Then for every $\omega = (x_1, x_2, ..., x_{100}) \in \Omega$, we could define X as:

$$X(\omega) = (x_1, x_2) \tag{1}$$

and

$$h(x) = 1_{x_2 \geqslant x_1}(x) = \begin{cases} 1, if x_2 \geqslant x_1 \\ 0, otherwise \end{cases} \tag{2}$$

In that case, X is a much more succinct description of the system than Ω, yet constraints 1 and 2 are still satisfied since the CDF of X is known explicitly (it is the product of the CDF of each die since they are assumed independent) and the metric of interest, p is simply:

$$p = P(X_2 \geqslant X_1) = E[1_{x_2 \geqslant x_1}(X)] = E[h(X)]$$

As a side note, a more succinct representation of the system than that in equation 1 is $X(\omega) = x_1 - x_2$. It can readily be seen that this representation meets constraint 1 above by taking $h(x) = 1_{x \leqslant 0}(x)$ but whether it meets constraint 2 hinges on whether we can obtain a closed form expression for

the CDF $X_1 - X_2$. In this particular case, it can be obtained by a simple convolution; and in fact the entire problem can be solved analytically or numerically. We will often use analytically tractable problems like this one as benchmarks of the performance of the simulation we develop.

So far we have only defined the simulation framework and shown how problems of interest can be expressed within it. We will now examine ways of solving efficiently, through simulation, a problem expressed in this framework.

2.2 Direct Monte-Carlo simulation

Given a simulation problem formulated in the framework of section 2.1, a straightforward method to obtain an estimate for the quantity of interest, $p = E[h(X)]$, is to simulate the system independently n times and average the results:

$$\hat{p}_{MC}^{n} = \frac{1}{n}\sum_{i=1}^{n} h(x_i) \tag{3}$$

In equation 3, the x_i are vectors drawn independently from the distribution of X and $\hat{p}_{MC}^{n}$ is an unbiased estimator of the solution to the problem obtained using the direct Monte-Carlo method. It follows from the weak law of large number that in the limit as n goes towards infinity, equation 3 converges in probability to p, or equivalently, $\lim_{n\to\infty} P(|\hat{p}_{MC}^{n} - p| > \varepsilon) = 0$ for any $\varepsilon > 0$.

In any practical simulation, n will actually be finite so we need to quantify the trade-off between computational cost (which grows with increasing

n) and error (which decreases, in a probabilistic sense to be defined below, with n). In order to do so, we will require that $p > 0$ and that the second moment of $h(X)$, $E\left[h(X)^2\right]$ exist, so that we can build confidence intervals for $\hat{p}_{MC}^n$ using the standard Gaussian approximation suggested by the Central Limit Theorem. With $\sigma = \sqrt{E\left[h(X)^2\right] - E\left[h(X)\right]^2}$, we write that an integer n_0 induces an (α, β) confidence interval around p if:

$$P\left(\frac{|\hat{p}_{MC}^{n_0} - p|}{p} < \beta\right) \geqslant \alpha \tag{4}$$

where $\alpha \in [0,1]$ and $\beta > 0$. For example, a (0.95,0.01) confidence interval is induced when n_0 is large enough that the probability that $\hat{p}_{MC}^{n_0}$ is within 1% of p is greater than 95%. Approximating $\hat{p}_{MC}^{n_0}$ with a Gaussian random variable with mean p and variance $\frac{\sigma^2}{n_0}$ and isolating n_0 in equation 4, we obtain[3]:

$$n_0 \geqslant \left(\frac{\sigma}{p}\right)^2 \left(\frac{\varphi^{-1}\left(\frac{\alpha+1}{2}\right)}{\beta}\right)^2 \tag{5}$$

where $\varphi^{-1}(\cdot)$ is the inverse function of the CDF of a standard normal random variable:

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{t^2}{2}} dt \tag{6}$$

Equation 5 illustrates the trade-off between computational cost and accuracy of the Monte-Carlo method. It reveals that, for a given confidence interval requirement, the number of iterations required is proportional to the variance of the random variable $h(X)$. For example, in the (0.95,0.01) confidence interval example, equation 5 implies that $n_0 > 38,415\left(\frac{\sigma}{p}\right)^2$ (see

[3]See appendix A for more details.

table 3, page 73), namely that we need to take and average at least 38,415 times the squared coefficient of variation of $h(X)$ many samples to induce a (0.95,0.01) confidence interval on $\hat{p}_{MC}^{n_0}$.

Equation 5 suggests that the performance of a Monte-Carlo estimator for a given quantity of interest p is captured by the variance σ^2 of the underlying $h(X)$ in our simulation framework. This means that among multiple estimators for the same quantity p, the best estimator is the one with the smallest value of σ in the sense that it requires the least computational effort to meet a given confidence interval[4]. Or alternatively, for a given amount of computational resources, it provides the tightest confidence interval. Therefore from now on, we will focus on finding estimators that make σ in equation 5 as small as possible. One way of reducing the variance of an estimator is by using the method of importance sampling, as illustrated in the next section.

2.3 Importance Sampling

In section 2.1 we introduced a simulation framework in which simulation problems can be expressed as a quadruplet: a probability sample space Ω, a random vector X defined on Ω, the CDF F_X of X, and a function $h(\cdot)$. The solution to a problem expressed in this framework is a real number $p = E[h(X)]$. In section 2.2, we have shown a classical algorithm, Monte-Carlo simulation, that can be used to obtain an arbitrarily close estimate of the solution, $\hat{p}_{MC}^{n}$, to a problem expressed in our simulation framework. Furthermore we have shown that the computational cost of the Monte-Carlo

[4]It is worth mentioning that the computational cost also depends on the cost of obtaining each sample $h(x_i)$. For the purpose of this work, we assume that this cost is finite and identical for all estimators.

method is directly proportional to the variance of the random variable $h(X)$. In this section, we introduce the method of importance sampling.

Definition 2.2. *A map ζ from a set of simulation problems Ξ to another set Ξ' is called solution-invariant if, for every simulation problems $(\Omega, X, F_X, h) \in \Xi$ and $\left(\tilde{\Omega}, \tilde{X}, \tilde{F}_{\tilde{X}}, \tilde{h}\right) \in \Xi'$ such that $\left(\tilde{\Omega}, \tilde{X}, \tilde{F}_{\tilde{X}}, \tilde{h}\right) = \zeta\left((\Omega, X, F_X, h)\right)$,*

$$E_{F_X}\left[h\left(X\right)\right] = E_{\tilde{F}_{\tilde{X}}}\left[\tilde{h}\left(\tilde{X}\right)\right] \tag{7}$$

where the notation $E_{F_X}\left[h\left(X\right)\right]$ stands for the expected value of $h(X)$ under the CDF F_X, namely:

$$E_{F_X}\left[h(X)\right] = \int_{\mathbb{R}^n} h\left(X\right) dF_X$$

Definition 2.2 hints at what we are trying to do. That is to say, to find a way to map a simulation problem into another simulation problem that yields an estimator with a lower variance of $h(X)$, and use the Monte-Carlo technique of section 2.2 to solve the latter problem instead. If our map is solution-invariant, then the estimate of the transformed solution is also an estimate of the solution of our initial problem.

Definition 2.3. *A map ζ from a set of simulation problems Ξ to another set Ξ' is called an importance sampling change of measure if for every measurable set $A \subseteq \mathbb{R}^n$, and every simulation problems $(\Omega, X, F_X, h) \in \Xi$ and $\left(\tilde{\Omega}, \tilde{X}, \tilde{F}_{\tilde{X}}, \tilde{h}\right) \in \Xi'$ such that $\left(\tilde{\Omega}, \tilde{X}, \tilde{F}_{\tilde{X}}, \tilde{h}\right) = \zeta\left((\Omega, X, F_X, h)\right)$,*

$$\int_A h(X) dF_X > 0 \Rightarrow \int_A d\tilde{F}_{\tilde{X}} > 0 \tag{8}$$

and,

$$\tilde{h}\left(\tilde{X}(\omega)\right) \overset{a.s.}{=} h(X(\omega)) \frac{dF_X}{d\tilde{F}_{\tilde{X}}}(\omega) \tag{9}$$

where the notation $\overset{a.s.}{=}$ *means that the set on which the equation does not hold has* dF_X *measure 0, and* $\frac{dF_X}{d\tilde{F}_{\tilde{X}}}(\omega)$ *is the Radon-Nikodym derivative*[5] *of* F_X *with respect to* $\tilde{F}_{\tilde{X}}$ *evaluated at* ω.

Note that in the case where X and $\tilde{X}$ are continuous random vectors with probability density functions $f_X(x)$ and $f_{\tilde{X}}(x)$ respectively, $\frac{dF_X}{d\tilde{F}_{\tilde{X}}}(\omega)$ reduces to $\frac{f_X(X(\omega))}{f_{\tilde{X}}(X(\omega))}$ so that equation 9 becomes:

$$\tilde{h}\left(\tilde{X}(\omega)\right) \overset{a.s.}{=} h(X(\omega)) \frac{f_X(X(\omega))}{f_{\tilde{X}}(X(\omega))} \tag{10}$$

The term $\frac{f_X(X(\omega))}{f_{\tilde{X}}(X(\omega))}$ is the likelihood ratio of observing a sample $X(\omega)$ under probability density function f_X relative to that of observing it under $f_{\tilde{X}}$.

Theorem 2.1. *Every importance sampling change of measure is a solution-invariant map.*

Proof. Suppose ζ is an importance sampling change of measure from Ξ to Ξ', and let $(\Omega, X, F_X, h) \in \Xi$ and $\left(\tilde{\Omega}, \tilde{X}, \tilde{F}_{\tilde{X}}, \tilde{h}\right) = \zeta((\Omega, X, F_X, h))$, then:

$$E_{\tilde{F}_{\tilde{X}}}\left[\tilde{h}\left(\tilde{X}\right)\right] = \int_{\mathbb{R}^n} \tilde{h}\left(\tilde{X}\right) d\tilde{F}_{\tilde{X}} = \int_{\mathbb{R}^n} h(X) \frac{dF_X}{d\tilde{F}_{\tilde{X}}} d\tilde{F}_{\tilde{X}} = \int_{\mathbb{R}^n} h(X)\, dF_X$$

$$= E_{F_X}[h(X)]$$

[5] See [14] for a treatment of the Radon-Nikodym derivative.

where the second equality comes from the definition of an importance sampling change of measure (eqn. 9), and the third is an application of the Radon-Nikodym theorem.

□

Theorem 2.1 tells us that we can change the CDF function F_X to any other CDF $\tilde{F}_{\tilde{X}}$ (subject to equation 8) and after weighing $h(X)$ by $\frac{dF_X}{d\tilde{F}_{\tilde{X}}}$, the solution to the simulation problem remains unchanged. Importance sampling in our simulation framework is thus a mapping from a simulation problem $(X, F_X(x), h(x))$ to another simulation problem $\left(\tilde{X}, F_{\tilde{X}}(x), \tilde{h}(x)\right)$ that has the same solution.

Note however that the cost of running the Monte-Carlo algorithm on $\left(\tilde{\Omega}, \tilde{X}, \tilde{F}_{\tilde{X}}, \tilde{h}\right)$ need not be the same as that of running it on (Ω, X, F_X, h) even though both simulation problems have the same solution. As we have shown in section 2.2, the computational cost of solving a simulation problem is directly proportional to the variance of $h(X)$. Thus, when faced with a problem (Ω, X, F_X, h) our objective is to find a new measure $\tilde{F}_{\tilde{X}}$ whose importance sampling change of measure yields a lower variance, or equivalently since importance sampling is solution-invariant, a lower second moment:

$$\int_{\mathbb{R}^n} \tilde{h}\left(\tilde{X}\right)^2 d\tilde{F}_{\tilde{X}} < \int_{\mathbb{R}^n} h(X)^2 \, dF_X$$

by using equation 9, this reduces to:

$$\int_{\mathbb{R}^n} h(X)^2 \left(\frac{dF_X}{d\tilde{F}_{\tilde{X}}}(X)\right)^2 d\tilde{F}_{\tilde{X}} < \int_{\mathbb{R}^n} h(X)^2 \, dF_X \tag{11}$$

There is no known practical method to find an optimal measure $\tilde{F}_{\tilde{X}}$ in the sense of minimizing the left hand side of equation 11 in the general case. However, a common simulation problem, a subset of which this thesis's contribution addresses, is one where $h(X)$ is an indicator function:

$$h(X) = 1_A(X)$$

for some measurable set $A \subset \mathbb{R}^n$. This case is treated in the next section.

2.4 Rare Events

In the case of a simulation problem where $h(X) = 1_A(X)$ for some measurable set $A \subset \mathbb{R}^n$, solving the problem corresponds to evaluating the probability that the vector X takes a value inside the set A:

$$p = E[1_A(X)] = P(X \in A)$$

This is often useful in models of networks when we wish to determine the likelihood that the system enters a bad state A. For example if X is a vector of queue-lengths, and $A = \left\{(x_1, x_2, ..., x_n) \mid \sum_i x_i > N\right\}$, p corresponds to the probability that the number of customers in the system exceeds a threshold N by some specified time T. As for any simulation problem, we can resort to the direct Monte-Carlo method described in section 2.2 to solve such a problem[6]. In that case, evaluating equation 3 for $\hat{p}^n_{MC}$ is equivalent

[6]Generally, we can only apply direct Monte-Carlo simulation to a system that we can simulate in finite time (see footnote 4 on page 14) which excludes direct simulation of the steady-state probability of a given event so T must be finite. We show in section 3 however that it is possible to exploit analytical properties of some queuing system to

to simulating the system n times, and producing the estimate:

$$\hat{p}_{MC}^{n} = \frac{k}{n}$$

where k is the number of times, in the n trials, that the system entered the state A. Notice that the distribution of $h(X) = 1_A(X)$ is Bernouilli and therefore has variance $\sigma^2 = p(1-p)$. Thus equation 5 suggests that the minimum number of trials needed to achieve a given (α, β) confidence interval must be at least:

$$n_0 \geqslant K\frac{(1-p)}{p} \tag{12}$$

where $K = \left(\frac{\varphi^{-1}\left(\frac{\alpha+1}{2}\right)}{\beta}\right)^2$ is fixed for a given confidence interval specification. So long as the probability of the system entering the set A is large enough, direct Monte-Carlo simulation provides a viable way of estimating it. Note that when p is small (as is the case in many system reliability, or insurance models) the number of replications required becomes prohibitive very quickly. As a quantitative example, using the (0.95,1.01) confidence interval of section 2.2, simulating a system with a failure probability of $p = 10^{-9}$ would require at least 3.9×10^{13} replications, which is unpractical even with modern hardware. The purpose of the remainder of this section is to illustrate how importance sampling can be used to reduce, in some cases by several orders of magnitude, the number of iterations required to achieve a specified confidence interval.

Definition 2.4. *A simulation problem* (Ω, X, F_X, h) *is called an event prob-*

find a solution-invariant map from the steady-state problem ($T = \infty$) to one that can be simulated in finite time.

lem if there exists a dF_X-measurable set $A \subseteq \mathbb{R}^n$ such that:

$$h(X) = 1_A(X)$$

where $1_A(X)$ is an indicator function that takes value 1 when $X \in A$ and 0 otherwise.

In the case of an event problem, with $h(X) = 1_A(X)$, let $\left(\tilde{\Omega}, \tilde{X}, \tilde{F}_{\tilde{X}}, \tilde{h}\right)$ be a candidate transformed problem under an importance sampling change of measure. The performance of the Monte-Carlo algorithm on this problem is dictated by the second moment of the estimator $\tilde{h}\left(\tilde{X}\right)$ (left hand side of equation 11) which reduces to:

$$\int_{\mathbb{R}^n} \tilde{h}\left(\tilde{X}\right)^2 d\tilde{F}_{\tilde{X}} = \int_{\mathbb{R}^n} h(X)^2 \left(\frac{dF_X}{d\tilde{F}_{\tilde{X}}}(X)\right)^2 d\tilde{F}_{\tilde{X}} = \int_{\mathbb{R}^n} 1_A(X)^2 \left(\frac{dF_X}{d\tilde{F}_{\tilde{X}}}(X)\right)^2 d\tilde{F}_{\tilde{X}}$$
$$= \int_A \left(\frac{dF_X}{d\tilde{F}_{\tilde{X}}}(X)\right)^2 d\tilde{F}_{\tilde{X}} \tag{13}$$

Since the computational cost of applying the Monte-Carlo algorithm on problem $\left(\tilde{\Omega}, \tilde{X}, \tilde{F}_{\tilde{X}}, \tilde{h}\right)$ increases with the second moment of $\tilde{h}\left(\tilde{X}\right)$, finding the least computationally expensive importance sampling measure coincides with finding the measure $\tilde{F}_{\tilde{X}}$ that minimizes $\frac{dF_X}{d\tilde{F}_{\tilde{X}}}(X)$ on the set A according to equation 13.

Theorem 2.2. *Fix Ω and let Ξ_r be the set of event problems defined on Ω and Ξ the set of simulation problems also defined on Ω. There exists a*

unique importance sampling change of measure $\zeta^* : \Xi_r \to \Xi$ *such that*

$$\text{var}\left(h^*\left(X^*\right)\right) = 0$$

whenever $(\Omega, X^*, F_X^*, h^*) = \zeta^*\left((\Omega, X, F_X, h)\right)$.

Proof. Let $(\Omega, X, F_X, h) \in \Xi_r$. By definition 2.4, there exists a measurable set A such that:

$$h\left(X\right) = 1_A\left(X\right)$$

Define ζ^* by:

$$\zeta^*\left((\Omega, X, F_X, 1_A)\right) = \left(\Omega, X, F_X^*, \frac{1_A}{p}\right) \tag{14}$$

where $p = E_{F_X}\left[1_A\right] = \int_{\mathbb{R}^n} h\left(X\right) dF_X$ is the solution to the problem and F_X^* is given by:

$$\begin{aligned} F_X^*\left(x\right) &= \int_Q \frac{1_A\left(z_1, z_2, ..., z_n\right)}{p} dF_X \\ Q &= \left\{\left(z_1, z_2, ..., z_n\right) | z_1 \leqslant x_1, z_2 \leqslant x_2, ..., z_n \leqslant x_n\right\} \end{aligned} \tag{15}$$

The second moment of h^*, using equation 13 is given by:

$$\int_{\mathbb{R}^n} h^*\left(X\right)^2 dF_X = \int_A \left(\frac{dF_X}{dF_X^*}\left(X\right)\right)^2 dF^* = \int_A p^2 dF^* = p^2$$

where the second inequality comes from the fact that $\frac{dF^*}{dF} = \frac{1_A}{p}$. Finally,

$$\text{var}\left(h^*\left(X\right)\right) = E\left[h^*\left(X\right)^2\right] - \left(E\left[h^*\left(X\right)\right]\right)^2 = p^2 - p^2 = 0 \tag{16}$$

□

Corollary 2.3. *For every solution-invariant map* $\zeta : \Xi \to \Xi_r$, *and for every problems* $(\Omega, X, F_X, h) \in \Xi$ *and* $(\Omega, X', F'_X, h') = \zeta((\Omega, X, F_X, h))$,

$$E_{F'_X}\left[h'\left(X'\right)^2\right] \geqslant E_{F^*_X}\left[h^*\left(X^*\right)^2\right]$$

Proof. From positivity of variance and equation 16,

$$\begin{aligned} &\operatorname{var}\left(h'\left(X\right)\right) \geqslant \operatorname{var}\left(h^*\left(X\right)\right) \\ &E\left[h'\left(X\right)^2\right] - \left(E\left[h'\left(X\right)\right]\right)^2 \geqslant E\left[h^*\left(X\right)^2\right] - \left(E\left[h^*\left(X\right)\right]\right)^2 \end{aligned} \tag{17}$$

But since both ζ and ζ^* are solution-invariant maps, it follows from equation 7 that:

$$E_{F'_{X'}}\left[h'\left(X\right)\right] = E_{F^*_{X^*}}\left[h^*\left(X\right)\right] = E_{F_X}\left[h\left(X\right)\right] \tag{18}$$

Substituting equation 18 in equation 17 yields the desired result. □

Theorem 2.2 is a remarkable result. It not only suggests that there is a 0-variance estimator for the solution of any rare-event problem but it also gives us a formula (see equation 15) to compute it explicitly.

To illustrate this with an example, suppose we would like to solve the following problem:

$$\left(\Omega = \mathbb{R}, X\left(\omega\right) = \omega, F_X\left(x\right) = \varphi\left(x\right), h\left(X\right) = 1_{x \in [2, 2.5]}\right) \tag{19}$$

where $\varphi\left(\cdot\right)$ is the CDF of a standard normal Gaussian random variable (see 6).

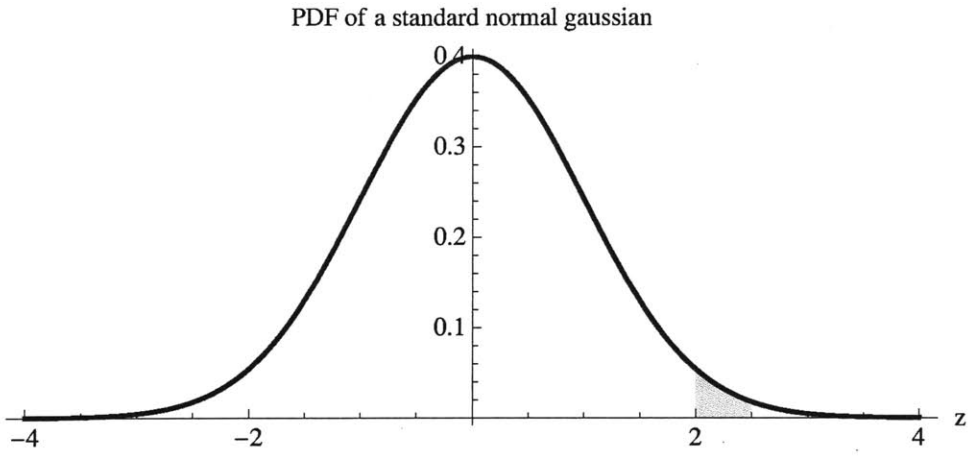

Figure 2: Graphical illustration of the example problem 19. Solving it is equivalent to computing the shaded area.

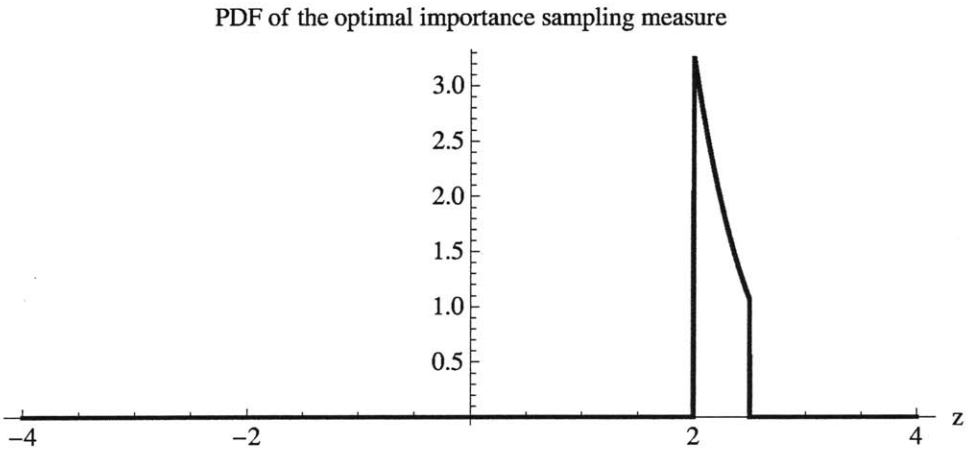

Figure 3: Graphical illustration of the PDF of the optimal change of measure for problem 19 as given by theorem 2.2, equation 15.

Figures 2 and 3 illustrate graphically the problem and the optimal sampling distribution given by theorem 15. Notice that the optimal distribution (figure 3) puts the entire probability mass on the event $A = [2, 2.5]$, the rare event of problem 19.

The difficulty in using the result of theorem 15 in practice is that the expression for the optimal measure (equation 15) depends on p, the very quantity we are trying to estimate. If p was known, there would be no need to use importance sampling to estimate it. Nevertheless, theorem 15 and its corollary 2.3 hint at the characteristics that a good (in the sense of having low variance) importance sampling measure should possess. Namely, it should put a large probability mass on the rare-event.

This is expected since equation 13 already hinted that the second moment of $\tilde{h}\left(\tilde{X}\right)$ is smallest when the term $\left(\frac{dF_X}{d\tilde{F}_{\tilde{X}}}(X)\right)^2$ takes small values on the rare-event A. Recall from equation 10 that in the case of a continuous random vector X, $\frac{dF_X}{d\tilde{F}_{\tilde{X}}}(X)$ is simply the likelihood ratio so that minimizing it over A is akin to maximizing the denominator, $\tilde{f}(x)$, on the set A.

The standard approach, and indeed the one we will use, is, given a simulation problem, to parametrize the set of candidate solutions and find the parameter (analytically or numerically) that minimizes the variance of the important sampling estimator. So far we have discussed the form of the optimal importance sampling distribution as well as given an intuitive criterion for what constitutes a good importance sampling change of measure. In the next section, we give a formal performance criterion for importance sampling change of measures.

2.5 Asymptotic optimality

Asymptotic optimality is a commonly-used criterion in the literature to evaluate importance sampling changes of measure. As we will refer to it several times throughout this thesis, we introduce it here in the context of our simulation framework.

Definition 2.5. *A rare-event problem set is a set of event problems Ξ with a common probability space Ω, a common random variable X, and a common distribution function F_X, together with a bijection r from the interval $(0, p_{\max})$ for some $p_{\max} > 0$ to Ξ such that for every $p \in (0, p_{\max})$, the rare-event problem $\left(\Omega, X, F_X, 1_{A_p}\right) = r(p)$ has solution p, namely:*

$$\int_{A_p} dF_X = p$$

Definition 2.5 can be thought of as a set of dF_X-measurable sets A_p indexed by their probability. The only criterion we require is that no matter how small a positive probability p we want, there exists a problem with an event set of smaller measure. We do not require the sets A_p to be nested when p decreases although this will be the case in all the examples we will look at. We can now define asymptotic optimality:

Definition 2.6. *An importance sampling change of measure ζ from a rare-event problem set Ξ to another set of problems Ξ' is asymptotically optimal if and only if:*

$$\liminf_{p \to 0} \frac{\log\left(\int_{\mathbb{R}^n} \tilde{h}_p^2 d\tilde{F}_{\tilde{X}}\right)}{\log\left(\int_{\mathbb{R}^n} \tilde{h}_p d\tilde{F}_{\tilde{X}}\right)} \geqslant 2 \tag{20}$$

where $\tilde{h}_p$ *and* $d\tilde{F}_{\tilde{X}}$ *are the estimator and the measure respectively of the problem that* ζ *maps* $r(p)$ *to. That is to say:*

$$\left(\Omega, \tilde{X}, \tilde{F}_{\tilde{X}}, \tilde{h}_p\right) = \zeta\left(r\left(p\right)\right)$$

Notice that, by positivity of the variance of $\tilde{h}_p\left(\tilde{X}\right)$:

$$\int_{\mathbb{R}^n} \tilde{h}_p^2 d\tilde{F}_{\tilde{X}} \geqslant \left(\int_{\mathbb{R}^n} \tilde{h}_p d\tilde{F}_{\tilde{X}}\right)^2$$

$$\log\left(\int_{\mathbb{R}^n} \tilde{h}_p^2 d\tilde{F}_{\tilde{X}}\right) \geqslant 2\log\left(\int_{\mathbb{R}^n} \tilde{h}_p d\tilde{F}_{\tilde{X}}\right)$$

$$\frac{\log\left(\int_{\mathbb{R}^n} \tilde{h}_p^2 d\tilde{F}_{\tilde{X}}\right)}{\log\left(\int_{\mathbb{R}^n} \tilde{h}_p d\tilde{F}_{\tilde{X}}\right)} \leqslant 2$$

$$\limsup_{p\to 0} \frac{\log\left(\int_{\mathbb{R}^n} \tilde{h}_p^2 d\tilde{F}_{\tilde{X}}\right)}{\log\left(\int_{\mathbb{R}^n} \tilde{h}_p d\tilde{F}_{\tilde{X}}\right)} \leqslant 2$$

so that equation 20 is equivalent to:

$$\lim_{p\to 0} \frac{\log\left(\int_{\mathbb{R}^n} \tilde{h}_p^2 d\tilde{F}_{\tilde{X}}\right)}{\log\left(\int_{\mathbb{R}^n} \tilde{h}_p d\tilde{F}_{\tilde{X}}\right)} = 2$$

We begin with a negative result to illustrate definition 2.6:

Theorem 2.4. *The identity map* $\zeta_i : \Xi_r \to \Xi_r$ *is not asymptotically optimal.*

Proof. Since $\zeta_i : \Xi_r \to \Xi_r$ is the identity map, we have $\left(\Omega, X, F_X, 1_{A_p}\right) = \zeta_i\left(\left(\Omega, X, F_X, 1_{A_p}\right)\right)$ so that we may substitute $\tilde{h}_p = 1_{A_p}$ and $d\tilde{F}_{\tilde{X}} = dF_X$ in

the left-hand side of equation 20:

$$\liminf_{p\to 0} \frac{\log\left(\int_{\mathbb{R}^n} \tilde{h}_p^2 d\tilde{F}_{\tilde{X}}\right)}{\log\left(\int_{\mathbb{R}^n} \tilde{h}_p d\tilde{F}_{\tilde{X}}\right)} = \liminf_{p\to 0} \frac{\log\left(\int_{\mathbb{R}^n} 1_{A_p}^2 dF_X\right)}{\log\left(\int_{\mathbb{R}^n} 1_{A_p} dF_X\right)} = \liminf_{p\to 0} \frac{\log(p)}{\log(p)} = 1 < 2$$

Thus ζ_i is not asymptotically optimal. □

Thus, no rare-event problem (in the sense of definition 2.4) is asymptotically optimal as stated. Instead, it has to be transformed by an aptly selected importance sampling change of measure. The optimal change of measure, ζ^*, from theorem 2.2 is an example of such a change of measure, as illustrated below:

Theorem 2.5. *The map ζ^* defined by equations 14 and 15 is asymptotically optimal.*

Proof. With $(\Omega, X, F_X^*, h^*) = \zeta^*(r(p))$, theorem 2.2 yields:

$$\begin{aligned}
&\operatorname{var}(h^*(X)) = 0 \\
&\int_{\mathbb{R}^n} (h^*)^2 d\tilde{F}_{\tilde{X}} = \left(\int_{\mathbb{R}^n} h^* d\tilde{F}_{\tilde{X}}\right)^2 \\
&\frac{\log\left(\int_{\mathbb{R}^n} (h^*)^2 d\tilde{F}_{\tilde{X}}\right)}{\log\left(\int_{\mathbb{R}^n} h^* d\tilde{F}_{\tilde{X}}\right)} = 2
\end{aligned}$$

From which equation 20 immediately follows. □

This result should not be surprising as the importance sampling change

of measure ζ^* was shown to have the lowest variance and second moment among all importance sampling changes of measure.

In this section, we have established a rigorous simulation framework within which we can compute performance metrics. We have examined the direct Monte-Carlo method and have introduced the variance reduction method of importance sampling. Our framework is generic enough to be applicable to a wide range range of simulation problems. In the next section, we restrict our attention to simulating networking problems, culminating with the presentation of the method of hazard rate twisting from [11], upon which this thesis's contribution builds.

3 Network Simulations

In this section, we examine how to apply importance sampling to network simulation problems. For our purpose, network simulation problems are simulation problems consisting of a set of queues linked by a group of nodes (we will use the terms node and server interchangeably). In particular, we will focus exclusively on networks with queues that have i.i.d. (independent and identically distributed) inter-arrival and service times.

We will begin by examining single queues in isolation and we will touch upon importance sampling methods in the particular case where they have light-tailed service time distributions. We will then turn to networks of queues, highlighting the difficulties encountered in applying single queue methods in the context of non-trivial networks. We will finally look at adaptive importance sampling, and how it can address some of the difficulties encountered in applying importance sampling simulation methods to queues, culminating with the main contribution of this thesis: the introduction of the adaptive version of the delayed hazard tate twisting algorithm[7].

3.1 Steady-state distributions

Let $Q(t)$ be a non-negative integer valued stochastic process describing the evolution of the length of a single-server queue subject to the boundary condition $Q(0) = 0$ and i.i.d. inter-arrival times $A_1, A_2, ...$ and service times $B_1, B_2, ...$ with specified distributions F_A and F_B respectively (see figure 4).

[7]The original hazard rate twisting algorithm was first introduced in [11] by Juneja and Shahabuddin.

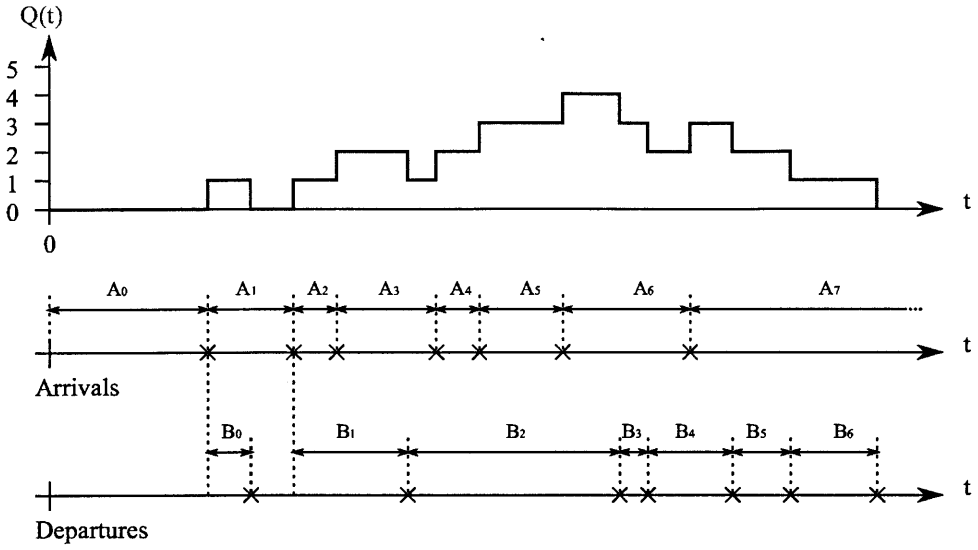

Figure 4: An example sample path for $Q(t)$.

Definition 3.1. *The inter-arrival times $A_0, A_1, \ldots$ and the service times $B_0, B_1, \ldots$ of a queue are called admissible if they are independent and identically distributed, and satisfy:*

$$E[A_0] > E[B_0] > 0$$

Theorem 3.1. *If $Q(t)$ is the queue-length process of a single-server queue with admissible inter-arrival times and service times, and W_m is the waiting time of the m^{th} customer, then there exist a cumulative distribution function $W(t)$ such that for every $t > 0$:*

$$\lim_{m\to\infty} P(W_m \leqslant t) = W(t) \tag{21}$$

Proof. See [7]. □

While theorem 3.1 guarantees the convergence in distribution of the waiting time, it gives no indication of how to compute $W(t)$. It is worth mentioning that in the particular case where the queue is M/M/1, that is to say A_i and B_j both are i.i.d. and exponentially distributed with $E[A_i] = \lambda^{-1} > \mu^{-1} = E[B_j]$, it is well known that:

Theorem 3.2 (Waiting time distribution in an M/M/1 queue). *If A_i and B_j both are i.i.d. and exponentially distributed with $E[A_i] = \lambda^{-1} > \mu^{-1} = E[B_j]$, then:*

$$W(t) = (1-\rho)\left(1 + \sum_{i=1}^{\infty} \rho^i \frac{\gamma(i, \mu t)}{i - 1!}\right) \tag{22}$$

where $\rho = \frac{\lambda}{\mu} < 1$ and $\gamma(k, z)$ is the lower incomplete gamma function.

Proof. Let Q be a random variable with the steady-state distribution of the queue length of an M/M/1 queue, it is well known [4] that $P(Q = n) = (1-\rho)\rho^n$, thus:

$$\begin{aligned}
W(t) &= \lim_{m\to\infty} P(W_m \leqslant t) \\
&= \lim_{m\to\infty} \sum_{i=0}^{\infty} P(W_m \leqslant t, Q_m = i) \\
&= \lim_{m\to\infty} \sum_{i=0}^{\infty} P(W_m \leqslant t | Q_m = i)\, P(Q_m = i) \\
&= \lim_{m\to\infty} P(W_m \leqslant t | Q_m = 0)\, P(Q_m = 0) \\
&+ \sum_{i=1}^{\infty} P(W_m \leqslant t | Q_m = i)\, P(Q_m = i) \\
&= 1 \cdot (1-\rho) + \lim_{m\to\infty} (1-\rho) \sum_{i=1}^{\infty} \rho^i P(W_m \leqslant t | Q_m = i) \\
&= (1-\rho)\left(1 + \sum_{i=1}^{\infty} \rho^i \frac{\gamma(i, \mu t)}{i - 1!}\right)
\end{aligned}$$

where Q_m is the queue length when the m^{th} customer enters the queue. The last equality comes from the fact that, conditioned on the queue length being i when the m^{th} customer arrives, the waiting time of that customer is the sum of i exponential random variables with rate μ, which has an erlang distribution with CDF $\frac{\gamma(i,\mu t)}{i-1!}$. □

Unfortunately however, there is no closed-form expression for $W(t)$ for arbitrary inter-arrival and service time distributions so that we must often resort to simulation in order to characterize $W(t)$. Two problems arise in attempting to simulate $W(t)$:

1. Estimating $W(t)$ for all possible values of t using simulation is not possible because there are infinitely many values of t. Thus we must restrict our objective to estimating finitely many quantities that characterize $W(t)$ within the simulation framework of section 2.1.

2. The quantity $W(t)$ is defined in terms of a limit as $m \to \infty$ (equation 21) which does not make it immediately clear how one could estimate any numerical quantity that describes $W(t)$ in a finite amount of time.

With regards to the first problem, we will restrict our attention to characterizing the probability that the waiting time exceeds some specified T, namely estimating $\overline{W(T)} = 1 - W(T)$, the steady-state probability that a customer waits longer than T. The reason for that is that if we estimate $\overline{W(T)}$ for several values T, we can use these estimates to form an empirical estimate of the CDF $\widehat{W}(T)$. We can in turn apply statistical techniques such as the Hill estimator or the ratio estimator [13] to this empirical estimate to obtain insight of queuing delays.

As far as the second problem is concerned, there are generally two approaches that can be taken. The first is to upper bound the difference (relative to some aptly chosen metric) between $P(W_m \leqslant t)$ and $W(t)$. If our upper-bound is tight enough, we would expect it to converge rapidly to 0 as m increases. We could then use a union-bound type argument to find a m large enough (but finite) for which we could simulate the system such that the total error between our estimate $\widehat{W}(t)$ and the true value is no greater than the sum of the errors between $P(W_m \leqslant t)$ and $W(t)$, and that between our estimate $\widehat{W}(t)$ and $P(W_m \leqslant t)$. The second approach, and indeed the one we will use, consists of transforming the simulation problem into another simulation problem that can be simulated in finite time and whose solution is exactly $W(t)$. The remainder of this chapter is dedicated to doing just that.

3.2 Importance sampling and GI/GI/1 queues

We begin by establishing an equivalence between the steady-state probability of a customer entering a GI/GI/1 queue waiting longer than t and that of a random walk exceeding a that threshold. We then show how an appropriate choice of an importance sampling estimator makes the simulation time required to obtain a single sample almost surely finite, paving the way for the application of the tools discussed in chapter 2.

Theorem 3.3 (Lindley's recursion). *For every non-negative interger n, let*

W_n be the time that the n^{th} customer spends in the queue. Then,

$$\begin{aligned} W_0 &= 0 \\ W_{n+1} &= (W_n + X_{n+1})^+ \end{aligned} \tag{23}$$

where $(\cdot)^+ = \max(\cdot, 0)$ and

$$X_{n+1} = B_n - A_{n+1} \tag{24}$$

Proof. First, observe that the first customer finds an idle server so it never spends any time in queue:

$$W_0 = 0$$

Also, when $A_{n+1} > W_n + B_n$, customer n has departed before customer $n+1$ arrives so the system is idle and in that case, $W_{n+1} = 0$. Lastly, when $A_{n+1} \leqslant W_n + B_n$, we have (see figure 5):

$$W_n + B_n = A_{n+1} + W_{n+1}$$

Solving for W_{n+1} in the case where when $A_{n+1} \leqslant W_n + B_n$ thus trivially yields:

$$W_{n+1} = W_n + B_n - A_{n+1}$$

So that for both cases, we may summarize the evolution of W_n by the equation:

$$W_{n+1} = (W_n + X_{n+1})^+$$

with $X_{n+1} = B_n - A_{n+1}$, as claimed. □

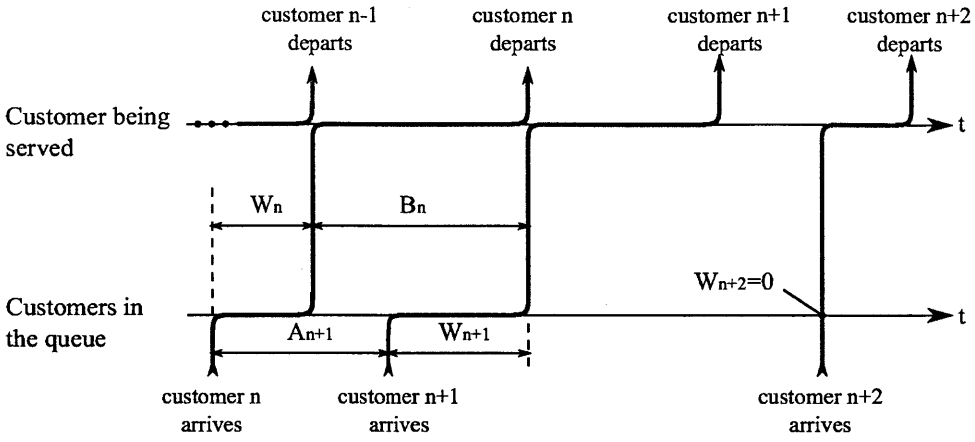

Figure 5: Diagram illustrating Lindley's recursion (theorem 3.3)

Corollary 3.4 (Random walk equivalence).

$$W_n = \max\{S_0^n, S_1^n, ..., S_n^n\}$$

where $S_0^n = 0$ *and* $S_i^n = \sum_{j=n-i+1}^{n} X_j$ *for* $i > 0$.

Proof. We will prove this by recurrence on n. The case $n = 0$ is trivial. Suppose that $W_n = \max\{S_0^n, S_1^n, ..., S_n^n\}$. Then, by the update equation 23, we have:

$$\begin{aligned}
W_{n+1} &= \max(W_n + X_{n+1}, 0) \\
&= \max(\max\{S_0^n, S_1^n, ..., S_n^n\} + X_{n+1}, 0) \\
&= \max(\max\{S_0^n + X_{n+1}, S_1^n + X_{n+1}, ..., S_n^n + X_{n+1}\}, 0) \\
&= \max\{0, S_0^n + X_{n+1}, S_1^n + X_{n+1}, ..., S_n^n + X_{n+1}\} \\
&= \max\{S_0^{n+1}, S_1^{n+1}, S_2^{n+1}, ..., S_{n+1}^{n+1}\}
\end{aligned}$$

where the last equality stems from the fact that $S_i^n + X_{n+1} = S_{i+1}^{n+1}$. Thus the claim follows by induction. □

Corollary 3.4 is a powerful result, one that will serve us more than theorem 3.3. Indeed, observe that the sequence $S_0^n, S_1^n, S_2^n, ..., S_n^n$ defines a random walk starting at 0, with i.i.d. increments $X_n, X_{n-1}, ..., X_0$. In particular, when the inter-arrival time and service time distributions are admissible, the increments have strictly negative mean. Thus corollary 3.4 reveals that W_n is the maximum of the first n terms of a random walk with negative drift. Thus the steady-state distribution of the waiting times can be obtained by estimating the probability that the maximum of a random walk with negative drift exceeds some threshold. Specifically,

$$\overline{W(t)} = \lim_{m\to\infty} P(W_m > t) = \lim_{m\to\infty} P(\max\{S_0^m, S_1^m, ..., S_m^m\} > n) \qquad (25)$$

However, estimating such a probability to any confidence interval in our simulation framework is not directly feasible because it would require us to draw an infinite number of samples S_i^m with positive probability. That is due to the fact that the random walk $S_0^m, S_1^m, S_2^m, ..., S_m^m$ has negative mean and thus has a positive probability of never reaching n. Importance sampling can alleviate that problem if we are able to find an alternative measure under which the random walk has positive drift since positive drift random walks almost surely exceed any threshold in finite time [9]. Exponential twisting, the subject of the remainder of this chapter, is such a change of measure for certain class of GI/GI/1 queues.

Before embarking on an exposition of importance sampling, we sim-

plify the notation from chapter 2 somewhat. While the explicit notation (Ω, X, F_X, h) served us well to rigorously establish the main results of chapter 2, it is somewhat cumbersome. In this chapter, we restrict our attention to continuous random variables for which there exists a probability density function f. We denote the solution to a simulation problem as $E_f[h(X)]$ where the subscript f in the expectation operator indicates that the expectation is taken over the original distribution. Unless otherwise specified, the sample space is implicitly $\Omega = \mathbb{R}^n$ and X is an identity map over Ω. Given an importance sampling change of measure, we denote the new probability density function as $\tilde{f}$. With this notation, theorem 2.1, in the case of continuous random variables, would be stated as:

$$E_f[h(X)] = E_{\tilde{f}}[h(X)\,L(X)] \tag{26}$$

where $L(X)$ is the likelihood ratio of equation 10, i.e., $L(X) = \frac{f(X)}{\tilde{f}(X)}$.

Let τ_n be the index at which the random walk with i.i.d. increments X_i first exceeds some threshold t (in the case where the random walk always takes value less than or equal to n, we have $\tau_n = \infty$). We can then rewrite equation 25 as:

$$\overline{W(t)} = E_f\left[1_{\{\tau_n < \infty\}}\right] \tag{27}$$

where f is the joint probability distribution of the i.i.d. increments X_i, i.e.,

$$f(x_1, x_2, ..., x_m) = \prod_{i=1}^{m} f(x_i) = \prod_{i=1}^{m} \left[\int_{-\infty}^{\infty} f_A(\zeta)\, f_B(x_i + \zeta)\, d\xi\right]$$

with f_A and f_B being the probability density of the inter-arrival time and

service time respectively and the integral representing the probability distribution of $X_i = B_{i-1} - A_i$ from equation 24.

Lemma 3.5. *Let X be a continuous random variable with negative expectation and moment generating function $M_X(\theta) = E\left[e^{\theta X}\right]$. If there exists an open interval including 0 on which $M_X(\theta)$ is finite, then there exists a unique positive θ^* such that:*

$$M_X(\theta^*) = 1$$

Furthermore,

$$\frac{d}{d\theta} M_X(\theta)\Big|_{\theta=\theta^*} > 0$$

Proof. Because the expected value of X is negative, the derivative of $M_X(\theta)$ at 0 must be negative:

$$\frac{d}{d\theta} M_X(\theta)\Big|_{\theta=0} = E[X] < 0$$

Thus, there must exist some $\theta_0 > 0$ such that $M_X(\theta_0) < 1$ since $M_X(0) = 1$. Furthermore, since moment-generating functions are not bounded above on $\mathbb{R}^+$ [8], there must exist some $\theta_2 > \theta_0$ such that $M_X(\theta_2) > 1$. Since $M_X(\theta)$ is a continuous function, there must therefore also exist a θ^* such that $\theta^* \in (\theta_0, \theta_2)$ and

$$M_X(\theta^*) = 1$$

The uniqueness of θ^* stems from the convexity of moment-generating functions on the neighborhood of 0 on which they are finite. Thus it follows that

$M_X(\theta^*)$ is increasing at θ^*, therefore:

$$\frac{d}{d\theta} M_X(\theta)\Big|_{\theta=\theta^*} > 0$$

□

Theorem 3.6. *Let X be a continuous random variable with negative mean, probability density function $f_X(x)$ and moment generating function $M_X(\theta)$. If there exists an open interval containing 0 where $M_X(\theta)$ is finite, then the random variable $\tilde{X}$ with probability density function $\tilde{f}_X = e^{\theta^* x} f_X(x)$, where when θ^* is the positive solution of the equation $M_X(\theta) = 1$, has positive mean.*

Proof. Let $M_{\tilde{X}}(\theta)$ be the moment generating function of $\tilde{X}$. Then:

$$\begin{aligned}
E\left[\tilde{X}\right] &= \frac{d}{d\theta}\left[M_{\tilde{X}}(\theta)\right]\Big|_{\theta=0} \\
&= \frac{d}{d\theta}\left[\int_{-\infty}^{\infty} e^{\theta x} e^{\theta^* x} f_X(x)\, dx\right]\Big|_{\theta=0} \\
&= \frac{d}{d\theta}\left[M_X(\theta+\theta^*)\right]\Big|_{\theta=0} \\
&= \frac{d}{d\theta'}\left[M_X(\theta')\right]\Big|_{\theta'=0} > 0
\end{aligned}$$

where the third equality comes from a change of variable $\theta' = \theta + \theta^*$ and the inequality is a consequence of lemma 3.5. □

Corollary 3.7. *If a random walk with i.i.d. increments $X_1, X_2, \ldots$ has negative drift and that the moment generating function of X_i converges in some open interval around 0, then the random walk with i.i.d. increments*

$\tilde{X}_1, \tilde{X}_2, \ldots$ *where* $\tilde{X}_i$ *has probability density* $\tilde{f}_X = e^{\theta^* x} f_X(x)$ *has positive drift.*

Proof. Direct application of theorem 3.6. □

Returning to equation 27, in the case where X has a moment generating function that is finite on some open interval B around 0, let us define a family of alternative candidate measures for X by:

$$f_\theta(x) = \frac{e^{\theta x} f_X(x)}{M_X(\theta)} \tag{28}$$

for every $\theta \in B$. Then, by theorem 2.1 and equation 27, we find that:

$$\overline{W(t)} = E_{f_\theta}\left[1_{\{\tau_n < \infty\}} L_\theta(X)\right] \tag{29}$$

where $L_\theta = \frac{f(X)}{f_\theta(X)} = \prod_{i=1}^{\tau_n} \frac{f(X_i)}{f_\theta(X_i)}$ is the likelihood ratio. In particular, when $\theta = \theta^*$, by corollary 3.7, the random walk with increments with density given by equation 28 has positive drift, so $\{\tau_n < \infty\}$ almost surely, and equation 29 becomes:

$$\overline{W(t)} = E_{f_{\theta^*}}\left[\prod_{i=1}^{\tau_n} \frac{M_X(\theta^*)}{e^{\theta^* X_i}}\right] = E_{f_{\theta^*}}\left[\exp\left(-\theta^* \sum_{i=1}^{\tau_n} X_i\right)\right] \tag{30}$$

Equation 30 in effect gives us a recipe for computing the probability that the waiting time of a customer entering a $G/G/1$ queue for some threshold n. Assuming we can bound the variance of the likelihood ratio, the steps are as follow:

1. Compute the probability density function, f_X, of X by convoluting

that of the interval arrival times, f_A with that of the service times, f_B:

$$f_X(x) = \int_{-\infty}^{\infty} f_A(\zeta) f_B(x+\zeta) d\xi$$

2. Compute the moment generating function $M_X(\theta)$ of X:

$$M_X(\theta) = \int_{-\infty}^{\infty} f_X(x) e^{\theta x} dx$$

3. If there exists an open interval around 0 on which $M_X(\theta)$ exists, then lemma 3.5 guarantees that the equation

$$M_X(\theta) = 1$$

has a unique positive root. Solve for this root (perhaps numerically) and label it θ^*.

4. Draw i.i.d. samples $x_1, x_2, ...$ from the following distribution:

$$f_{\theta^*}(x) = f_X(x) e^{\theta^* x}$$

until

$$\sum_{i=1}^{m} x_i > n$$

Notice that corollary 3.7 guarantees that this step will be completed in finite time.

5. Form an unbiased estimate of the probability that the waiting time

exceed t:

$$\widehat{W} = \exp\left(-\theta^* \sum_{i=1}^{m} x_i\right)$$

6. Repeat steps 4 and 5 as many times as needed to meet the desired confidence interval. The final waiting time probability estimate is an average of $\widehat{W}$ for all the trial runs.

So far, we have established a procedure, known in the literature as exponential twisting, for estimating the steady-state probability that a customer will experience a waiting time longer than t. We have turned a problem defined in equation 21 in terms of a limit as m goes towards infinity into another problem that yields an unbiased estimate of the desired quantity and that can be simulated in finite time. This is a remarkable result in its own right.

Nonetheless, we have not said anything about the performance of exponential twisting. It has been shown however in [16] that $f_{\theta^*}(x)$ is the only asymptotically optimal (in the sense of definition 2.6 when n goes towards infinity) measure in the class of measures $f_\theta(x)$. This result is further strengthened in [12] where it is shown that $f_{\theta^*}(x)$ is the optimal distribution in the class of all i.i.d. candidate distributions for the random walk's increment with positive mean. For an example of the method of exponential twisting as applied to the M/M/1 queue, see appendix C.

Notice that f_{θ^*} (and indeed the entire class of exponentially twisted distributions f_θ) only exists when the moment generating function of X converges in some open interval around 0. In light of corollary B.5, this restriction effectively precludes the application of the method of exponential

twisting to any queue where the service times follow a heavy-tailed distribution.

As was discussed in the introduction, systems with polynomially decaying service time distributions model empirically observed traces on data networks much more accurately than traditional exponentially decaying tail models [15]. This feature reflects the consistent empirical observation that a small fraction of users in public data networks consumes the overwhelming majority of the traffic. This in turns makes the ability to obtain large delay probabilities in such models especially valuable.

Unfortunately, to the best of our knowledge, there does not exist an asymptotically optimal importance sampling algorithm to simulate GI/GI/1 queues where the service time distribution's tail is allowed to decay polynomially. However, if we restrict the inter-arrival time distributions to an exponentially distributed one (i.e., an M/G/1 queue), there exists an approach to obtain the desired waiting time probabilities even when the service time distributions are heavy-tailed and that exponential twisting is not applicable. This is the subject of the following section.

3.3 The M/G/1 queue

In the previous section, our workhorse for converting the problem of estimating $W(t)$ in finite time was to use corollary 3.4 to turn a problem that, a priori, requires estimating a limit as m goes towards infinity into that of finding the probability that a random walk exceed a threshold. With an appropriate change of measure, we showed that the random walk could be made to have positive drift, guaranteeing it would exceed the specified

threshold, so that the expected value of the likelihood ratio gave us the desired waiting time probability.

In this section, we take a different approach to work around the problem of estimating a quantity in the limit. This approach rests on the following theorem:

Theorem 3.8 (The Pollaczeck-Khinchine formula). *Let $Q(t)$ describes the evolution of a queue with admissible inter-arrival and service time distributions with CDF F_A and F_B respectively. If F_A is the CDF of an exponential random variable with mean λ^{-1} (the M/G/1 case) and that B has mean μ_B, then:*

$$W(t) = (1-\rho)\sum_{k=0}^{\infty}\rho^k B_0^k(t)$$

where $\rho = \lambda\mu_B < 1$, B_0 is the residual time of B, i.e., $B_0(u) = \frac{1}{\mu_B}\int_0^u 1 - F_B(s)\,ds$ and B_0^k is the k^{th} convolution of B_0 with itself.

Proof. See [3]. □

Corollary 3.9.

$$W(t) = P(S_K \leqslant t)$$

where K is a geometric random variable with parameter ρ, $S_K = \sum_{i=1}^{K} X_i$, and X_i are i.i.d. random variable with probability density function $\frac{1}{\mu_B}(1 - F_B(x))$.

Proof.

$$\begin{aligned} P(S_K \leqslant t) &= P(S_K \leqslant t) \\ &= \sum_{k=0}^{\infty} P(K = k, S_k \leqslant t) \\ &= \sum_{k=0}^{\infty} P(K = k) P(S_k \leqslant t) \\ &= (1-\rho) \sum_{k=0}^{\infty} \rho^k B_0^k(t) \end{aligned}$$

□

Corollary 3.9 in effect gives us a recipe for estimating $W(t)$, one that was first shown in [2], namely repeat the following steps m times:

1. Generate a geometric random variable K with parameter ρ.
2. Generate $S_K = \sum_{i=1}^{K} X_i$ where the X_i have the distribution given in the corollary.

Then, an unbiased estimate of $W(t)$, $\widehat{W(t)}$ is given by $\widehat{W(t)} = \frac{k}{m}$ where k is the number of trials where $S_K > t$.

Notice that during the simulation, we do not need to know t. This means that once we have run the simulation, we can estimate $\widehat{W(t)}$ for any t very quickly, simply by counting the number of S_K that exceed t and dividing that by m. This process can be made especially efficient by sorting the outputs S_K of our simulation runs.

So far, we have not used importance sampling. In the remainder of this chapter, we will look at a particular class of importance sampling change of measure on X_i called delayed hazard rate twisting. We begin by establishing

a few definitions, and conclude with statements on the asymptotic optimality of delayed hazard rate twisting.

3.4 Delayed Hazard Rate Twisting

Definition 3.2. *A random variable X is called* long-tailed *if it is continuous, non-negative, and for every x, its CDF $F_X(x)$ satisfies:*

$$F_X(x) < 1$$

In light of definition 3.2, any random variable that cannot be upper-bounded almost surely by a finite constant is long-tailed. Most familiar non-degenerate continuous random variables (e.g. standard normal with positive variance, Poisson random variables with positive mean) fit this criterion.

Definition 3.3. *Given a long-tailed random variable X with CDF F_X and probability density function f_X, we denote by hazard rate $\lambda_X(x)$ and hazard function the function $\Lambda_X(x)$ the functions respectively defined by:*

$$\lambda_X(x) = \frac{f(x)}{1 - F(x)}$$
$$\Lambda_X(x) = \int_0^x \lambda_X(t)\, dt$$

Theorem 3.10. *Given a long-tailed random variable X with CDF F_X and hazard function $\Lambda_X(x)$,*

$$\Lambda_X(x) = -\log(1 - F_X(x))$$

Proof.

$$\begin{aligned}\Lambda_X(x) &= \int_0^x \lambda_X(t)\,dt \\ &= \int_0^x \frac{f_X(t)}{1-F_X(t)}dt \\ &= -\log(1-F_X(t))\Big|_{t=0}^{t=x} \\ &= -\log(1-F_X(x))\end{aligned}$$

□

Corollary 3.11. *Given a long-tailed random variable X with CDF F_X and hazard function $\Lambda_X(x)$,*

$$\lim_{x\to\infty} \Lambda_X(x) = \infty$$

Proof.

$$\lim_{x\to\infty} \Lambda_X(x) = \lim_{x\to\infty} -\log(1-F_X(x)) = \lim_{F_X\to 1} -\log(1-F_X) = \infty$$

□

Theorem 3.12. *Let X be a long-tailed random variable with CDF F_X, probability distribution function f_X, hazard rate function $\lambda(x)$ and hazard function $\Lambda(x)$. Then the function $f_X^\theta(x)$ defined in equation 31 for every $\theta \in [0,1)$ is a probability distribution function with associated hazard rate function $\lambda_X^\theta(x) = (1-\theta)\lambda(x)$ and CDF $F_X^\theta(x) = 1-(1-F(x))^{1-\theta}$.*

$$f_X^\theta(x) = \lambda(x)(1-\theta)\exp\{-(1-\theta)\Lambda(x)\} \tag{31}$$

Proof. First, we prove that $f_X^\theta(x)$ is a valid probability distribution function.

It follows immediately from equation 31 that $f_X^\theta(x) \geqslant 0$. Observe that:

$$\begin{aligned} f_X^\theta(x) &= \lambda(x)(1-\theta)\exp\{-(1-\theta)\Lambda(x)\} \\ &= \frac{f(x)(1-\theta)}{1-F(x)}\exp\left((1-\theta)\int_0^x \frac{f(t)}{1-F(t)}dt\right) \\ &= \frac{f(x)(1-\theta)}{1-F(x)}\exp\left((1-\theta)\log(1-F(x))\right) \\ &= \frac{f(x)(1-\theta)}{(1-F(x))^\theta} \end{aligned}$$

so that:

$$F_X^\theta(x) = \int_0^x f_X^\theta(t)\,dt = \int_0^x \frac{f(t)(1-\theta)}{(1-F(t))^\theta}dt = 1-(1-F(x))^{1-\theta}$$

It immediately follows that:

$$\int_0^\infty f_X^\theta(t)\,dt = \lim_{x\to\infty} 1-(1-F(x))^{1-\theta} = 1$$

so $f_X^\theta(x)$ is a valid probability distribution function with associated CDF $F_X^\theta(x) = 1-(1-F(x))^{1-\theta}$. Lastly,

$$\begin{aligned} \lambda_X^\theta(x) &= \frac{f_X^\theta(x)}{1-F_X^\theta(x)} \\ &= \frac{f(x)(1-\theta)}{\left(1-F_X^\theta(x)\right)^\theta\left(1-F_X^\Theta(x)\right)^{1-\theta}} \\ &= \frac{(1-\theta)f(x)}{1-F_X^\theta(x)} \\ &= (1-\theta)\lambda_X(x) \end{aligned}$$

□

Theorem 3.12 illustrates why the family of alternative importance sampling distributions $f_X^\theta(x)$ is known as hazard rate twisting. Indeed, it replaces f_X by a probability distribution f_X^θ whose hazard rate is a multiple of that of f_X. This family of importance sampling changes of measure is shown in [11] to be asymptotically optimal for estimating the probability that $P\left(\sum_{i=1}^{m} X_i > t\right)$ for some fixed number m when θ is appropriately chosen. It is also shown that the alternative distribution $f_X^{\theta,x_m}(x)$ given by equation 32 is asymptotically optimal for estimating $P\left(\sum_{i=1}^{M} X_i > t\right)$ when M is a geometrically distributed random variable for some x_m and θ. Recall from corollary 3.9 that $\overline{W(t)} = P\left(\sum_{i=1}^{M} X_i > t\right)$ when X_i have the same distribution as the residual time of the service time B.

$$f_X^{\theta,x_m}(x) = \begin{cases} f_X(x), x \leqslant x_m \\ \dfrac{1 - F_X(x_m)}{1 - F_X^\theta(x_m)} f_X^\theta(x), x > x_m \end{cases} \tag{32}$$

Unfortunately, the values of x_m and θ^* that make the estimator asymptotically optimal are known only in terms of the implicit solution of a family of equations involving the hazard function $\Lambda(x)$ of X. Furthermore, these equations are not guaranteed to have solutions for small values of n. While this is sufficient to establish asymptotic optimality, it provides little insight into choosing parameters that perform well in practical simulations. Our proposed algorithm, and indeed the contribution of this thesis, is an adaptive version of delayed hazard rate twisting that dynamically finds good parameters x_m and θ^* as the simulation is running. In the next section, we present this adaptive importance sampling algorithm, together with bounds

on its performance and comparisons to plain delayed hazard rate twisting, and exponential twisting in the light tailed case.

4 Adaptive Delayed Hazard Rate Twisting

In this final section, we present this thesis's contribution, the delayed hazard rate twisting algorithm, together with simulation results illustrating its performance in the special case of Pareto and Exponential[8] service time distributions in an M/G/1 queue.

We first start by presenting adaptive importance sampling in the single-variable casev (i.e., finding the optimal θ for the importance sampling change of measure f_X^θ in equation 31) and illustrate this with our adaptive plain hazard rate twisting algorithm. We then move on to adaptive importance sampling in the multi-variable case (i.e., finding optimal θ and x_m in f_X^{θ,x_m} in equation 32) and build on this to introduce the novel adaptive delayed hazard rate twisting algorithm. We conclude the chapter with simulation results showing the simulation performance gains of using the adaptive delayed hazard rate twisting algorithm, compared to direct Monte-Carlo.

4.1 Single-parameter adaptive importance sampling

Suppose we have an event-problem in the sense of definition 2.4 together with a parametrized family of candidate importance sampling changes of measure with likelihood ratio $L_\theta = \frac{f(X)}{f_\theta(X)}$. In this section, we will focus on the single-dimensional parameter case, that is to say we will assume that θ is a real number, possibly restricted to an interval, but X can be a vector of arbitrary dimensions.

[8]The waiting time distribution is well-known analytically for the M/M/1 queue (see equation 22), which makes it an especially attractive benchmark of our algorithm's performance.

Reusing the notation of equation 26, we can define an estimator Z_θ for the solution of the problem as:

$$Z_\theta = 1_A(X) L_\theta(X)$$

We will generally be interested in events A of the form $A = \left\{ \sum_{i=1}^{M} X_i > t \right\}$. Notice that it follows from theorem 2.1, for every θ_1 and θ_2,

$$P(X \in A) = E_{f_{\theta_1}}[Z_{\theta_1}] = E_{f_{\theta_2}}[Z_{\theta_2}] \tag{33}$$

However the variances $\text{var}[Z_{\theta_1}]$ and $\text{var}[Z_{\theta_2}]$ need not be the same. Indeed, in light of chapter 2, the best estimator within the class of estimators Z_θ is the one with the lowest variance. Thus our objective will be to find the value θ^* that minimizes the variance of Z_θ and then use that parameter in our simulation. More formally,

$$\theta^* = \arg\min_\theta \text{var}[Z_\theta] = \arg\min_\theta E_{f_\theta}\left[Z_\theta^2\right] - E_{f_\theta}[Z_\theta]^2 \tag{34}$$

Observe from equation 33 that $E_{f_\theta}[Z_\theta]^2$ is a constant term, so that equation 34 becomes:

$$\theta^* = \arg\min_\theta E_{f_\theta}\left[Z_\theta^2\right]$$

Using the notation

$$(\cdot)' = \frac{d(\cdot)}{d\theta}$$
$$(\cdot)^{(n)} = \frac{d^n(\cdot)}{d\theta^n}$$

and the likelihood ratio defined earlier by $L_\theta = \frac{f(X)}{f_\theta(X)}$ for every vector x and

real number θ, we then have the following theorem:

Theorem 4.1. *If $L_\theta(x)$ is a n-differentiable function of θ and that for every vector x, $\frac{\partial^n L_\theta}{\partial \theta^n}$ is bounded over the range of θ, then:*

$$v^{(n)}(\theta) = E_{f_X^\theta}\left[1_A(X) L_\theta(X) \frac{\partial^n L_\theta(X)}{\partial \theta^n}\right] \tag{35}$$

where $v(\theta) = E_{f_X^\theta}\left[Z_\theta^2\right]$.

Proof.

$$\begin{aligned}
v^{(n)}(\theta) &= \tfrac{d^n}{d\theta^n} E_{f_X^\theta}\left[Z_\theta^2\right] \\
&= \frac{d^n}{d\theta^n} E_{f_X^\theta}\left[1_A(X)^2 L_\theta(X)^2\right] \\
&= \frac{d^n}{d\theta^n} \int 1_A(x) L_\theta(x) L_\theta(x) f_X^\theta(x)\, dx \\
&= \frac{d^n}{d\theta^n} \int 1_A(x) L_\theta(x) f(x)\, dx \\
&= \int 1_A(x) f(x) \frac{\partial^n L_\theta(x)}{\partial x^n} dx \\
&= \int 1_A(x) L_\theta(x) \frac{\partial^n L_\theta(x)}{\partial x^n} f_X^\theta(x)\, dx \\
&= E_{f_X^\theta}\left[1_A(X) L_\theta(X) \frac{\partial^n L_\theta(X)}{\partial \theta^n}\right]
\end{aligned}$$

where the 4^{th} and 6^{th} equalities follow from observing that $f_X(x) = L_\theta(x) f_X^\theta(x)$ and the 5^{th} equality is a well-known result from real analysis that holds since $\frac{\partial^n L_\theta(x)}{\partial x^n}$ is bounded. □

Corollary 4.2. *If $v(\theta)$ is a convex function of θ and that there exists a θ_0 such that*

$$E_{f_X^{\theta_0}}\left[1_A(X) L_{\theta_0}(X) \left.\frac{\partial L_\theta(X)}{\partial \theta}\right|_{\theta=\theta_0}\right] = 0$$

then,

$$\theta^* = \theta_0$$

Proof. Since $v(\theta)$ is convex, it follows that $v'(\theta_0) = 0$ implies that $v(\theta_0)$ is a global minimum. But by theorem 4.1,

$$v'(\theta_0) = E_{f_X^{\theta_0}}\left[1_A(X) L_{\theta_0}(X) \left.\frac{\partial L_\theta(X)}{\partial \theta}\right|_{\theta=\theta_0}\right]$$

Thus the claim follows. □

Corollary 4.2 is an interesting side result in the case where $v(\theta)$ is convex, but is not necessary for the correctness of our algorithm. Observe that theorem 4.1 in effect provides us with a simulation problem in the form of chapter 2 which can be solved to obtain an estimate of any derivative of the second moment of our estimator as a function of the parameter θ. Furthermore note that all the derivatives $v^{(n)}(\theta)$ are the solutions of simulation problems that depend on the same distribution $f_X^\theta(x)$. This in turns means that we may estimate $v(\theta), v'(\theta), v^{(2)}(\theta)$ concurrently. Namely, we can draw a number of i.i.d. samples of X following the distribution f_X^θ and use these samples to estimate any derivative of $v(\theta)$ that we desire.

This opens the door to using classic gradient descent methods, including those that require access to the derivatives in order to minimize $v(\theta)$ (or equivalently, to find θ^*). Using Newton's method, we could then approximate θ^* in principle by updating repeatedly a guess $\widehat{\theta_0^*}$ with the following

update equation:

$$\widehat{\theta^*_{n+1}} = \widehat{\theta^*_n} - \delta \frac{\widehat{v'}\left(\widehat{\theta^*_n}\right)}{\widehat{v^{(2)}}\left(\widehat{\theta^*_n}\right)} \tag{36}$$

where $\delta > 0$ is a parameter that allows us to trade convergence rate versus accuracy and $\widehat{v'}$ and $\widehat{v^{(2)}}$ are the importance sampling estimates (using the simulation distribution $f_X^{\widehat{\theta^*_n}}$) of v' and $v^{(2)}$ obtained by using equation 35.

Corollary 4.2 further suggests that in cases where $v(\theta)$ is convex, it suffices to use a root-finding method on $v'(\theta)$ to find θ^*. In the next section, we propose an adaptive plain hazard rate twisting algorithm as a precursor to this thesis's main contribution: the adaptive delayed hazard rate algorithm.

4.2 Adaptive plain hazard rate twisting

Plain hazard rate twisting is an importance sampling change of measure f_X^θ illustrated by equation 31. It is proved in [11] that there exists some θ^* (that depends on m in the following equation) that makes this change of measure asymptotically optimal (as n goes towards infinity) for problems of the form:

$$P\left(\sum_{i=1}^{m} X_i > n\right)$$

where X_i are i.i.d. random variable with distribution f_X.

Observe in particular that this change of measure is known to be asymptotically optimal only for fixed m and thus it is not immediately useful for solving waiting times in M/G/1 queues of the forms given in theorem 3.8 which is our ultimate goal. It is worthwhile studying here however because it provides a stepping stone to the delayed hazard rate twisting change of

measure given in equation 32 which is known to be asymptotically optimal and ultimately to our novel adaptive delayed hazard rate twisting algorithm.

Evaluating the likelihood ratio expression L_θ for the plain hazard rate twisting change of measure yields:

$$\begin{aligned} L_\theta &= \frac{f_X(X)}{f_X^\theta(X)} \\ &= \prod_{i=1}^{m} \frac{f_X(X_i)}{f_X^\theta(X_i)} \\ &= \prod_{i=1}^{m} \frac{\lambda(X_i)\exp(-\Lambda(X_i))}{(1-\theta)\lambda(X_i)\exp(-(1-\theta)\Lambda(X_i))} \\ &= \prod_{i=1}^{m} (1-\theta)^{-1}\exp(-\theta\Lambda(X_i)) \\ &= (1-\theta)^{-m}\exp\left(-\theta\sum_{i=1}^{m}\Lambda(X_i)\right) \end{aligned} \tag{37}$$

It therefore follows that:

$$L_\theta^2 = (1-\theta)^{-2m}\exp\left(-2\theta\sum_{i=1}^{m}\Lambda(X_i)\right)$$

and the first and second partial derivatives of L_θ with respect to θ are therefore:

$$\begin{aligned} \frac{\partial L_\theta}{\partial\theta} &= \frac{\exp\left(-\theta\sum_{i=1}^{m}\Lambda(X_i)\right)\left(m+(\theta-1)\sum_{i=1}^{m}\Lambda(X_i)\right)}{(1-\theta)^{m+1}} \\ \frac{\partial^2 L_\theta}{\partial\theta^2} &= \frac{\exp\left(-\theta\sum_{i=1}^{m}\Lambda(X_i)\right)\left(m^2+m\left(1+2(\theta-1)\sum_{i=1}^{m}\Lambda(X_i)\right)+\left(\sum_{i=1}^{m}\Lambda(X_i)\right)^2(\theta-1)^2\right)}{(1-\theta)^{m+2}} \end{aligned} \tag{38}$$

Plugging equations 37 and 38 into equation 35 gives us a recipe for evaluating

$v(\theta)$, $v'(\theta)$ and $v^{(2)}(\theta)$ concurrently for various values of theta provided that we have an expression for $\Lambda(X_i)$, which can be computed using the result from theorem 3.10.

To illustrate the above, we will focus in the remainder of this section on the case where X_i is the residual time (see theorem 3.8) of a Pareto distributed random variable with minimum value $x_m > 0$ and scale parameter $\alpha > 1$. That is to say,

$$F_X(x) = \frac{1}{\mu_P} \int_0^x 1 - F_P(t)\, dt \tag{39}$$

with,

$$F_P(t) = \begin{cases} 0, t < x_m \\ 1 - \left(\frac{x_m}{t}\right)^{\alpha}, t \geqslant x_m \end{cases} \tag{40}$$

and

$$\mu_P = \frac{\alpha x_m}{\alpha - 1} \tag{41}$$

Substituting equation 40 into equation 39 and evaluating the integral gives:

$$F_X(x) = \begin{cases} 0, x < 0 \\ \left(\frac{\alpha - 1}{\alpha x_m}\right) x, 0 \leqslant x \leqslant x_m \\ \left(\frac{\alpha - 1}{\alpha x_m}\right) \cdot \left(x_m + \frac{x^{\alpha} x_m - x \cdot x_m^{\alpha}}{x^{\alpha}(\alpha - 1)}\right), x > x_m \end{cases} \tag{42}$$

Thus, by theorem 3.10 (page 46), $\Lambda_x(x) = -\log(1 - F_X(x))$ and substitut-

α	1.5
x_m	0.1667
n	100,000
m	10

Table 1: Simulation parameters

ing in equation 42 yields:

$$\Lambda_X(x) = \begin{cases} 0, x < 0 \\ \log\left(\frac{x_m\alpha}{x(1-\alpha)+x_m\alpha}\right), 0 \leqslant x \leqslant x_m \\ \log(\alpha) + (\alpha-1)\log\left(\frac{x}{x_m}\right), x > x_m \end{cases} \tag{43}$$

We now have all the pieces required to evaluate $P\left(\sum_{i=1}^{m} X_i > n\right)$ when X_i are i.i.d. random variables with the distribution of the residual time of a Pareto random variable. Substituting equation 43 into equation 38 and substituting the partial derivatives thus obtained into equation 35 gives us an implementable expression for computing $v(\theta)$ (the second moment of the estimator), $\frac{dv}{d\theta}$ and $\frac{d^2v}{d\theta^2}$.

As an illustration, we ran a simulation of $v(\theta)$ with the parameters given in table 1. Figures 6 through 9 illustrate the results of simulating the system for various values of θ, in practice $v(\theta)$, $v'(\theta)$ and $v^{(2)}(\theta)$ are only evaluated for the values of θ shown in figure 10. Observe how after only 4 steps (figure 10) our adaptive plain hazard rate twisting algorithm converges to θ^*. In this particular setup, our simulation gain is a reduction by a factor of more than 3 of simulation time (this corresponds to the ratio of the variance at

$\theta = 0$ to that of the variance at $\theta = \theta^*$ in figure 6). For the relationship between simulation time and variance of the estimator, see equation 53 in appendix A).

Figure 6 represents an unbiased estimate $E_{f_X^\theta}\left[1_A \prod_{i=1}^{m} L_\theta(X_i)\right]$ of $P\left(\sum_{i=1}^{m} X_i > t\right)$ with X_i being distributed like the residual time of a Pareto (42) and the parameters given in table 1. According to theorem 2.1, the actual solution is a straight line but it is not quite the case due to simulation noise. Figure 7 shows the variance of L_θ. Notice that at $\theta = 0$, the variance of the estimator is that of direct Monte-Carlo simulation since $f_X^{\theta=0}(x) = f_X(x)$. Observe that $\text{var}(L_{\theta=0}) = E[L_{\theta=0}](1 - E[L_{\theta=0}]) \approx E[L_{\theta=0}]$ thus as expected, figure this figure and figure 6 have approximately the same value at $\theta = 0$. It can be seen in this figure that $\theta^* \approx 0.3$. Figures 8 and 9 show respectively the derivative and second derivative of $v(\theta)$. Note that both were obtained directly using equation 35. Lastly figure 10 shows the convergence of the guess for θ^* (update equation 36).

Recall from the discussion in section 3.4 that plain hazard rate twisting is an asymptotically optimal importance sampling change of measure to compute $P\left(\sum_{i=1}^{m} X_i > t\right)$ for a fixed value of m, but not to compute $P\left(\sum_{i=1}^{M} X_i > t\right)$ when M is a geometric random variable, as in the Pollaczeck-Khinchine formula (theorem 3.8) which was our initial goal. Delayed hazard rate twisting however is asymptotically optimal in that case, but requires us to optimize two parameters simulataneously, θ and x_m in equation 32. In the remainder of this chapter, we do just that by presenting the adaptive delayed hazard rate twisting algorithm, the contribution of this thesis.

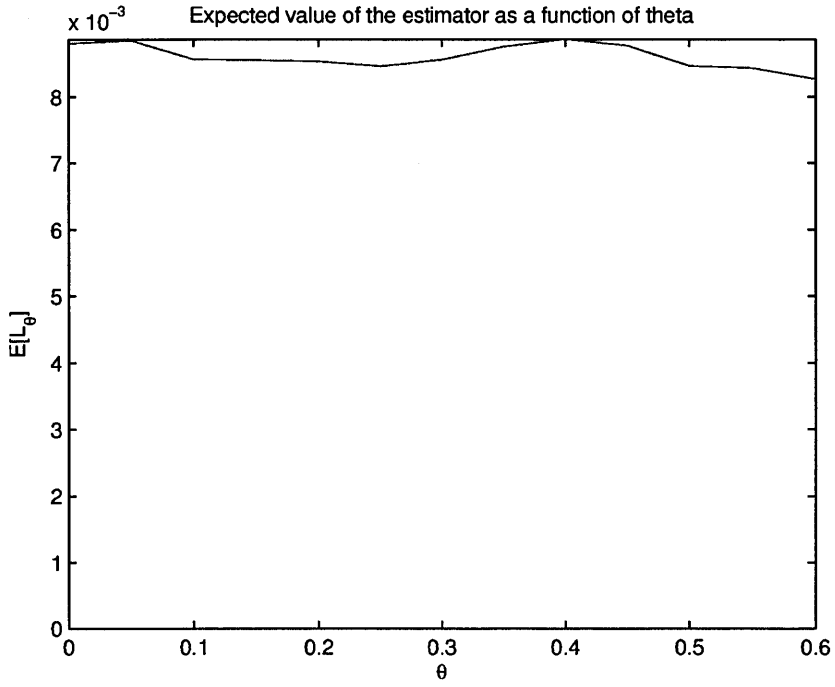

Figure 6: Plot of $E[L_\theta]$ as a function of θ.

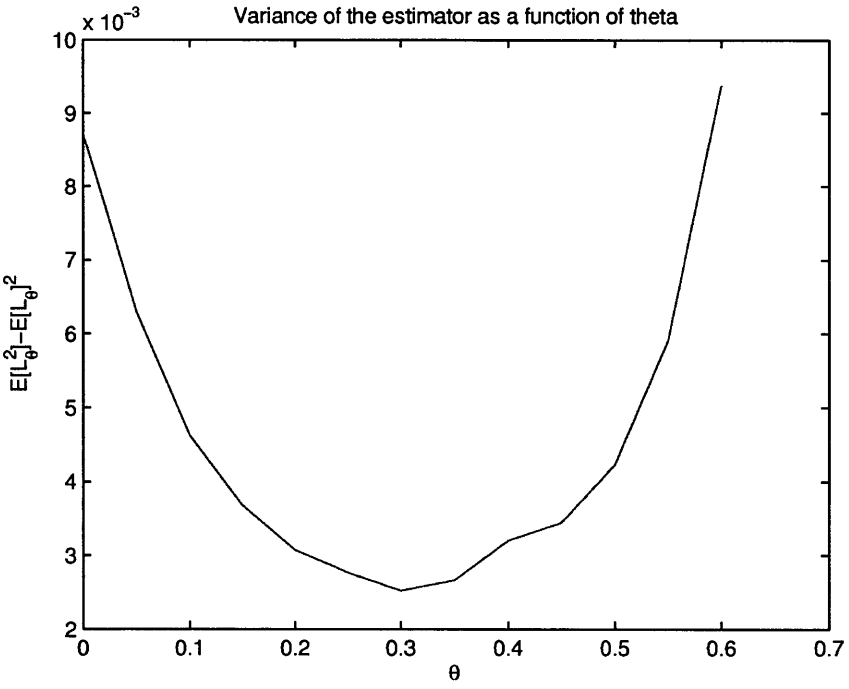

Figure 7: Variance of L_θ as a function of θ.

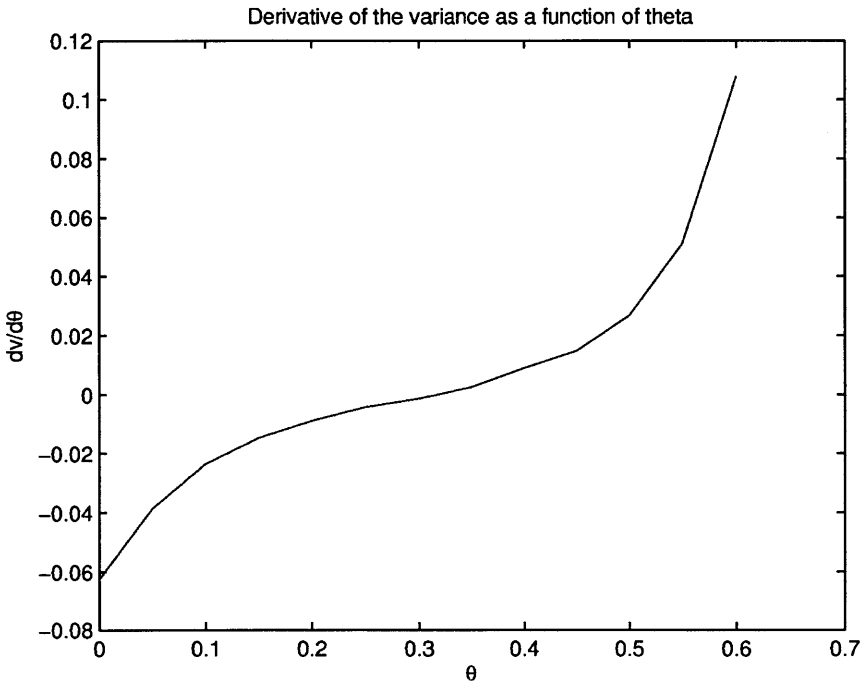

Figure 8: The derivative of the variance, $v'(\theta)$.

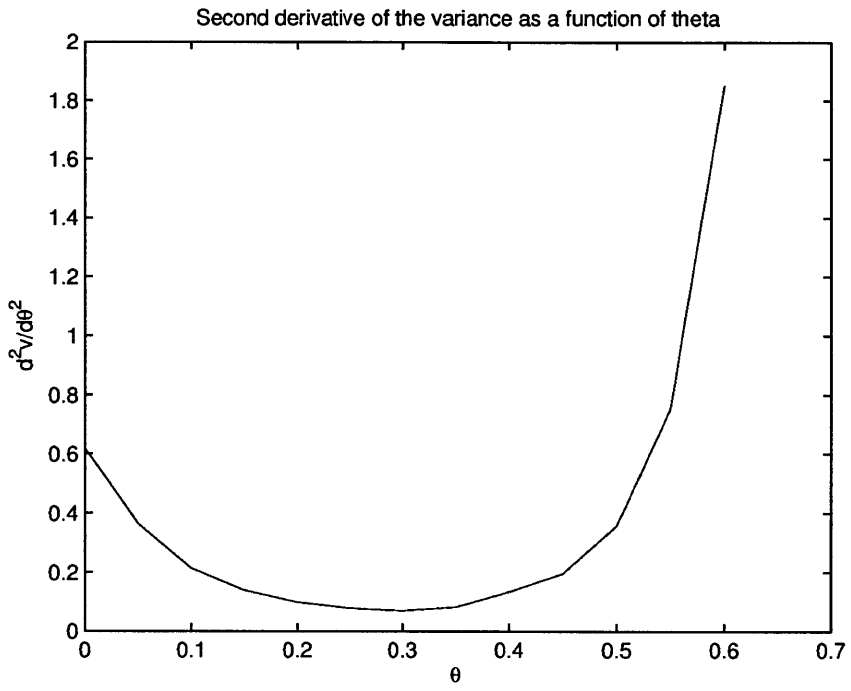

Figure 9: The second derivative of the variance, $v^{(2)}(\theta)$.

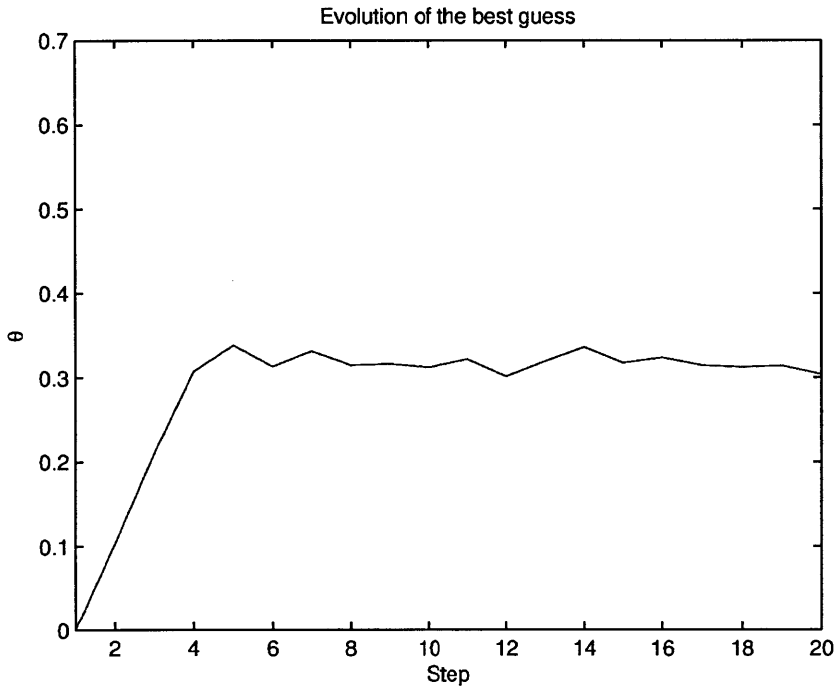

Figure 10: The adaptive plain hazard rate twisting algorithm root finding.

4.3 Multiple-parameter adaptive importance sampling

In the previous two sections, we looked at adaptive importance sampling when the family of candidate distributions is parametrized by a single parameter θ and that the likelihood ratio L_θ is a continuous function of that parameter. We then introduced the novel adaptive plain hazard rate twisting algorithm and illustrated it on the problem of estimating $P\left(\sum_{i=1}^{m} X_i > n\right)$ for a specific random variable X.

In this section, we look at the case where the family of candidate importance sampling changes of measure is parametrized by multiple parameters $\theta_1, \theta_2, ..., \theta_k$. In this case, we denote the candidate distribution by $f_X^{\boldsymbol{\theta}}(x)$ where $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_k)$ is a vector. Theorem 4.1 can easily be extended to the multiple parameter case to yield:

$$\frac{\partial v^{(n)}(\boldsymbol{\theta})}{\partial \theta_i} = E\left[1_A(X) L_{\boldsymbol{\theta}}(X) \frac{\partial^n L_{\boldsymbol{\theta}}(X)}{\partial \theta_i^n}\right] \tag{44}$$

for all $i \in \{1, 2, ..., k\}$ such that $\frac{\partial^n L_{\boldsymbol{\theta}}(X)}{\partial \theta_i^n}$ exists and is bounded for all X.

If $\frac{\partial^n L_{\boldsymbol{\theta}}(X)}{\partial \theta_i^n}$ exists and is bounded for all X, then we could use any of the optimization methods based on the Hessian matrix, H, of $v^{(n)}$ by iteratively evaluating H during our simulation and updating our guess of $\boldsymbol{\theta}^*$ using equation 44. This is not substantially different from the single-parameter case we studied in section 4.1.

Unfortunately, this approach does not work in the case where the likelihood ratio $L_{\boldsymbol{\theta}}(X)$ is not continuous with respect to all θ_i because equation 44 would not apply. The reason for that is explained below. Let $L_\theta^{x_h}$ be the likelihood ratio of f_X to $f_X^{\theta; x_h}$ (defined in equation 32). Then,

$$
\begin{aligned}
L_\theta^{x_h} &= \prod_{i=1}^{M} \frac{f_X(X_i)}{f_X^{\theta,x_h}(X_i)} \\
&= \prod_{X_i \leqslant x_h} 1 \prod_{X_i > x_h} \frac{1 - F_X(x_h)}{1 - F_X^\theta(x_h)} \frac{f_X(X_i)}{f_X^\theta} \\
&= \prod_{X_i > x_h} \frac{1 - F_X(x_h)}{1 - \left(1 - (1 - F_X(x_h))^{1-\theta}\right)} \frac{f_X(X_i)}{f_X^\theta} \\
&= \prod_{X_i > x_h} \frac{(1 - F_X(x_h))^\theta}{1 - \theta} \exp(-\theta \Lambda(X_i)) \\
&= \left(\frac{(1 - F_X(x_h))^\theta}{1 - \theta}\right)^{|\{X_i > x_h\}|} \exp\left(-\theta \sum_{X_i > x_h} \Lambda(X_i)\right)
\end{aligned}
\tag{45}
$$

In particular, observe that $\frac{\partial L_\theta^{x_h}}{\partial x_h}$ does not exist whenever $x_h = X_i$ for some i. Therefore we cannot apply theorem 4.1. In fact, we cannot use any optimization method that requires the partial derivate $\frac{\partial L_\theta^{x_h}}{\partial x_h}$. However the partial derivatives $\frac{\partial L_\theta^{x_h}}{\partial \theta}$ and $\frac{\partial^2 L_\theta^{x_h}}{\partial \theta^2}$ can be computed (albeit tediously) from equation 45, and are included below for completeness:

$$
\begin{aligned}
\frac{\partial L_\theta^{x_h}}{\partial \theta} &= \frac{\exp(-S\theta)\left(\frac{A^\theta}{1-\theta}\right)^Z \left(Z\left((1-\theta)\log A + 1\right) - S(1-\theta)\right)}{1-\theta} \\
\frac{\partial^2 L_\theta^{x_h}}{\partial \theta^2} &= \frac{\exp(-S\theta)\left(\frac{A^\theta}{1-\theta}\right)^Z \left(Z + (Z + S(\theta-1))^2 + Z(\theta-1)\log(A)\, C\right)}{(1-\theta)^2}
\end{aligned}
\tag{46}
$$

with

$$Z = |\{X_i | X_i > x_h\}|$$

$$A = 1 - F_X(x_h)$$

$$S = \sum_{\{X_i > x_h\}} \Lambda(X_i)$$

$$C = (-2(Z + S(-1 + \theta)) + Z(\theta - 1)\log(A))$$

Since the partial derivatives with respect to θ can be obtained directly, but not those with respect to x_h, we use a hybrid method as the optimization strategy for our adaptive delayed hazard rate twisting algorithm. This strategy is outlined in greater detail in the next and final section.

4.4 Adaptive delayed hazard rate twisting

We now extend the second moment of the estimator for the plain adaptive hazard rate twisting to the delayed adaptive hazard rate twisting case with a summation with a geometrically distributed number of terms:

$$v(\theta, x_h) = E_{f_X^{\theta,x_h}}\left[1_A \left(L_\theta^{x_h}\right)^2\right]$$

where $L_\theta^{x_h} = \prod_{i=1}^{M} \frac{f_X(X_i)}{f_X^{\theta,x_h}(X_i)}$ is the likelihood ratio of the distribution of X_i, f_X to the delayed hazard rate twisted distribution f_X^{θ,x_h} given in equation 32 and A is the event $A = \left\{\sum_{i=1}^{M} X_i > t\right\}$.

Recall from the previous chapter than when M is geometric with parameter $\rho = \lambda\mu_B$ (where λ is the arrival rate of the M/G/1 queue and μ_B is the mean service time) and that f_X is the distribution of the residual time of the service time distribution, then $E_{f_X^{\theta,x_h}}\left[1_A L_{\theta^*}^{x_h^*}\right]$ is equal to the probability

that a customer waits longer than t.

Thus $1_A L_\theta^{x_h}$ is an importance sampling estimator for $W(t)$ with second moment $E_{f_X^{\theta,x_h}}\left[1_A\left(L_\theta^{x_h}\right)^2\right] = v\left(\theta, x_h\right)$ so that finding the best (in the sense of lowest variance) estimator is akin to minimizing $v\left(\theta, x_h\right)$ with respect to θ and x_h and then using the values θ^* and x_h^* found to minimize $v\left(\theta, x_h\right)$ to compute $E_{f_X^{\theta,x_h}}\left[1_A\left(L_{\theta^*}^{x_h^*}\right)\right]$.

Since we can evaluate $\frac{\partial v(\theta,x_h)}{\partial \theta}$ (by substituting equations 45 and 46 into theorem 4.1) directly, but not $\frac{\partial v(\theta,x_h)}{\partial x_h}$, our approach consists of trying Q different candidate values for x_h^* and optimizing θ^* each time using the gradient descent outlined in equation 36. We choose the candidates solutions x_h^i as follows:

$$x_h^i = F_X^{-1}\left(\frac{i}{Q+1}\right)$$

for $i \in \{1, 2, ..., Q\}$.

Namely, for candidate solution x_h^i, a fraction approximately $i/(Q+1)$ of the samples will not be twisted, while the remainder will. For each i, we fix x_h^i and we find the value $\theta_{x_i}^*$ that minimizes $v\left(x_h^i, \theta_{x_i}^*\right)$ using the gradient descent. Finally, we select among all Q pairs $\left(x_h^i, \theta_{x_i}^*\right)$ the one that yields the smallest value of $v\left(x_h^i, \theta_{x_i}^*\right)$ and use that to simulate $E_{f_X^{\theta,x_h}}\left[1_A L_{\theta_{x_i}^*}^{x_h^i}\right]$.

Increasing Q leads to more granularity in the candidate $\left(x_h^i, \theta_{x_i}^*\right)$ at the expense of more simulation time, so Q acts as a trade-off between simulation cost and accuracy. Our simulations suggest that values of Q around 10 offer remarkable gains over plain adaptive hazard rate twisting. As a practical illustration of the performance gains achievable, we run our adaptive delayed hazard rate twisting algorithm and compare its performance to our adaptive

$\overline{W(t)}$	ADHZT	APHZT	DMC
$t = 10^1$	6×10^{-2} $(\pm 4.39 \times 10^{-3})$	5.48×10^{-1} $(\pm 4.89 \times 10^{-1})$	5.39×10^{-1} $(\pm 4.98 \times 10^{-1})$
$t = 10^2$	1.75×10^{-2} $(\pm 1.60 \times 10^{-3})$	2.39×10^{-1} $(\pm 4.05 \times 10^{-1})$	2.74×10^{-1} $(\pm 4.46 \times 10^{-1})$
$t = 10^3$	7.83×10^{-3} $(\pm 8.95 \times 10^{-4})$	8.5×10^{-2} $(\pm 2.42 \times 10^{-1})$	1.00×10^{-1} $(\pm 3 \times 10^{-1})$
$t = 10^4$	2.56×10^{-3} $(\pm 2.81 \times 10^{-4})$	2.42×10^{-2} $(\pm 1.27 \times 10^{-1})$	1.70×10^{-1} $(\pm 3.7 \times 10^{-2})$
$t = 10^5$	5.23×10^{-4} $(\pm 1.42 \times 10^{-4})$	9.01×10^{-3} $(\pm 7.13 \times 10^{-2})$	9.00×10^{-3} $(\pm 9.44 \times 10^{-2})$

Table 2: Comparison of the estimates of $\overline{W(t)} = 1 - W(t)$ as given by the adaptive delayed hazard rate twisting (ADHZT), the adaptive plain hazard rate twisting algorithm (APHZT) and direct Monte-Carlo simulation (DMC). The numbers in parenthesis are the standard deviations

plain hazard rate twisting algorithm and direct Monte-Carlo simulation for an M/G/1 queue with arrival rate $\lambda = 1.82$ and Pareto service time with $\alpha = 1.5$, $x_m = 0.1667$, $Q = 20$ and summarize our results in table 2.

Observe from table 2 that for large values of t, the standard deviations of the direct Monte-Carlo method and the adaptive plain hazard rate twisting algorithm are so high as to render them nearly useless relative to the mean values. The adaptive plain hazard rate twisting algorithm seems to perform somewhat better than the direct Monte-Carlo. The adaptive delayed hazard rate twisting algorithm however provides standard deviations two orders of magnitude smaller than either adaptive plain hazard rate twisting and direct Monte-Carlo in this particular setup for $t = 10^5$. This should not be very surprising because (non-adaptive) delayed hazard rate twisting is the only

algorithm of the three that is proven to be asymptotically optimal [11]. Also note that the probability of seeing a waiting time t appears to decrease subexponentially with t. This should also be expected since the service time distribution is heavy-tailed (with $\alpha = 1.5$, the Pareto distributed service time has finite mean, but infinite variance) in this setup.

5 Conclusion

Our main contribution in this thesis was an adaptive importance sampling algorithm for quickly estimating buffer overflow probabilities in stable M/G/1 queues. In particular, our algorithm operates even when the service times are heavy-tailed. We illustrated its operation on an M/G/1 queue with heavy-tailed Pareto service times with infinite variance, and showed simulation time gains of 2 orders of magnitude over naive simulation.

Perhaps one of the most interesting avenue for future work in the area of variance reduction for network simulations would be to develop a provably asymptotically optimal importance sampling change of measure for networks of 2 or more queues. The positive drift of the queue length with the asymptotically optimal importance sampling change of measure in the light-tailed G/G/1 queue case suggests that a promising avenue for estimators of buffer overflow probabilities for single queues and networks alike is to replace the original service time and arrival time distributions with those conditioned on the queue having overflown before returning to empty. If a sufficiently tight lower bound on the second moment of the estimator is obtained together with a sufficiently tight upper bound on its first moment, asymptotic optimality results would then follow.

A Confidence Intervals

This Appendix details the assumptions made in the construction of confidence intervals in section 2.2.

Given a sequence of n i.i.d. random variables $X_1, X_2, ..., X_n$ with variance σ^2 and mean μ, the Central Limit Theorem states that:

$$\lim_{n\to\infty} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma\sqrt{n}} \sim N(0,1) \tag{47}$$

where $\sim$ means that the left and right hand side have the same distributions, and $N(a,b)$ is a normal random variable with mean a and variance b. Notice that for all $n > 0$, the left hand side of 47 has mean 0 and variance 1. Our approximation relies on assuming that for n large enough,

$$\sum_{i=1}^{n} \frac{X_i - \mu}{\sigma\sqrt{n}} \sim N(0,1)$$

Multiplying both sides by $\frac{\sigma}{\sqrt{n}}$ and adding μ yields:

$$\frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right) \tag{48}$$

Observe that the left hand side of equation 48 is of the same form as equation 3. Continuing with this approximation, we may then approximate the distribution of $\hat{p}^n_{MC}$ by that of a normal with mean $\mu = E[h(X)] = p$ and a variance $s^2 = \frac{\sigma^2}{n} = \frac{E[h(X)^2] - E[h(X)]^2}{n}$:

$$\hat{p}^n_{MC} \sim N\left(p, \frac{E\left[h(X)^2\right] - E[h(X)]^2}{n}\right) \tag{49}$$

so that the probability that $\hat{p}^n_{MC}$ be within a fraction β from p is simply the probability α that a normal random variable with mean and variance given in equation 49 take values between $p(1-\beta)$ and $p(1+\beta)$. Namely:

$$\alpha = \frac{1}{s\sqrt{2\pi}} \int_{p(1-\beta)}^{p(1+\beta)} e^{-\frac{(t-p)^2}{2s^2}} dt \tag{50}$$

where s corresponds to the standard deviation of the normal random variable

$\alpha \downarrow \beta \rightarrow$	100%	50%	25%	10%	5%	1%
50%	4.55E-01	1.82E+00	7.28E+00	4.55E+01	1.82E+02	4.55E+03
75%	1.32E+00	5.29E+00	2.12E+01	1.32E+02	5.29E+02	1.32E+04
90%	2.71E+00	1.08E+01	4.33E+01	2.71E+02	1.08E+03	2.71E+04
95%	3.84E+00	1.54E+01	6.15E+01	3.84E+02	1.54E+03	3.84E+04
99%	6.63E+00	2.65E+01	1.06E+02	6.63E+02	2.65E+03	6.63E+04
99.9%	1.08E+01	4.33E+01	1.73E+02	1.08E+03	4.33E+03	1.08E+05

Table 3: The quantity $n/\left(\frac{\sigma}{p}\right)^2$ (equation 53) evaluated for select values of α and β.

in equation 49, i.e.,

$$s = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{E\left[h(X)^2\right] - E\left[h(X)\right]^2}{n}} \tag{51}$$

A simple change of variable in equation 50 (substitute $z = \frac{t-p}{s}$ in the integrand) yields:

$$\alpha = \varphi\left(\frac{p\beta}{s}\right) - \varphi\left(-\frac{p\beta}{s}\right) = 2\varphi\left(\frac{p\beta}{s}\right) - 1 \tag{52}$$

where $\varphi(\cdot)$ is defined in equation 6 (page 13) and the second equality follows from the identity $\varphi(x) = 1 - \varphi(-x)$. Equation 52 can be solved for s, yielding:

$$s = \frac{p\beta}{\varphi^{-1}\left(\frac{\alpha+1}{2}\right)}$$

where $\varphi^{-1}(\cdot)$ is the inverse of $\varphi(\cdot)$. Plots of both functions are given in figures 12 and 11 respectively. Substituting $s = \frac{\sigma}{\sqrt{n}}$ from equation 51 and solving for n yields the lower bound of equation 5 (page 13), i.e.:

$$n = \left(\frac{\sigma}{p}\right)^2 \left(\frac{\varphi^{-1}\left(\frac{\alpha+1}{2}\right)}{\beta}\right)^2 \tag{53}$$

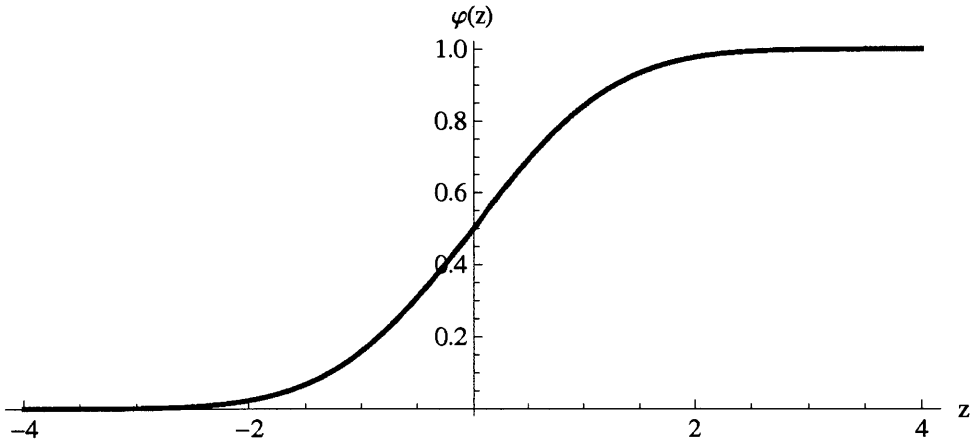

Figure 11: Plot of $\varphi(\cdot)$ defined in equation 6 (page 13).

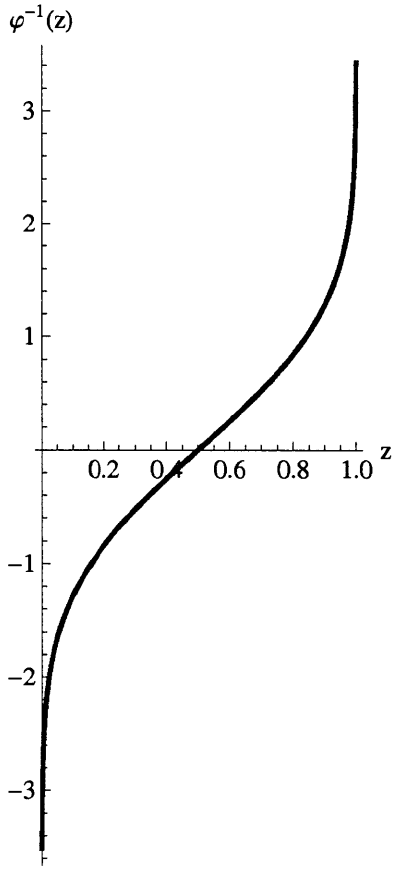

Figure 12: Plot of $\varphi^{-}1\,(\cdot)$.

B Heavy-tailedness

In this appendix, we define heavy-tailedness and collect several results that are useful for chapter 3.

Definition B.1. *The tail coefficient c of a non-negative random variable X is a quantity $c \in \mathbb{R}^+ \cup \{+\infty\}$ defined by:*

$$c = \sup_n \{n | E[X^n] < \infty \wedge n \geqslant 0\}$$

when the set $\{n|E[X^n] < \infty\}$ is not equal to $\mathbb{R}^+$. Otherwise we adopt the convention $c = \infty$.

Definition B.2. *A non-negative random variable X is called heavy-tailed if its tail coefficient is strictly less than 2.*

Lemma B.1. *Given a non-negative random variable X positive real number n, $E[X^n] = \infty$ if and only if $E[X^n | X^n > a] = \infty$ for every $a > 0$.*

Proof. By the total expectation theorem [5]:

$$E[X^n] = E[X^n | X^n \leqslant a] \cdot P(X \leqslant a) + E[X^n | X^n > a] \cdot P(X > a)$$

Since the term $E[X^n | X^n \leqslant a] \cdot P(X \leqslant a)$ is bounded above by a, the lemma follows immediately. □

Theorem B.2. *If the tail coefficient c of a non-negative random variable X is finite, then for every $d > c$,*

$$E\left[X^d\right] = \infty$$

Proof. Let X be a non-negative random variable with tail coefficient c. Let $d > c$. Then the function $f(x) = x^d - x^c$ is positive on the interval $(1, \infty)$. Thus:

$$\begin{aligned} &E[f(X) | X > 1] > 0 \\ &E\left[X^d - X^c | X > 1\right] > 0 \\ &E\left[X^d | X > 1\right] > E[X^c | X > 1] \\ &E\left[X^d\right] > E[X^c | X > 1] \cdot P(X > 1) + E\left[X^d | X \leqslant 1\right] \cdot P(X \leqslant 1) \end{aligned} \tag{54}$$

By lemma B.1, $E[X^c|X>1]=\infty$ thus the left hand side of 54, $E[X^d]$ must also be infinite since $E[X^d|X\leqslant 1]\cdot P(X\leqslant 1)$ is bounded below by 0. □

Corollary B.3. *Every heavy-tailed nonnegative random variable X has infinite variance.*

Proof. By definition B.2, if X is heavy tailed, then its tail coefficient c is less than 2. Applying theorem B.2 with $d=2$ yields that $E[X^2]=\infty$. Since $\operatorname{var}(X)=E[X^2]-E[X]^2$, then $\operatorname{var}(X)=\infty$. □

Theorem B.4. *Let X be a non-negative random variable with tail coefficient c and moment-generating function $M_X(s)=E[e^{sX}]$. If $c<\infty$, then there is no open-interval that includes 0 over which $M_X(s)$ exists is finite.*

Proof. It can be shown that for every $n>0$ and $s>0$,

$$\lim_{x\to\infty}\frac{e^{sx}}{x^n}=\infty \tag{55}$$

In particular, fix $\varepsilon>0$ and let $n=c+\varepsilon$, equation 55 implies that there exists an x_0 such that for every $x>x_0$,

$$e^{sx}>x^{c+\varepsilon}$$

or equivalently, the function $f(x)=e^{sx}-x^{c+\varepsilon}$ is strictly positive on (x_0,∞). Thus:

$$\begin{aligned}
&E[f(X)|X>x_0]>0\\
&E[e^{sX}|X>x_0]>E[X^{c+\varepsilon}|X>x_0]\\
&M_X(s)>E[X^{c+\varepsilon}|X>x_0]\,P(X>x_0)+E[e^{sX}|X\leqslant x_0]\,P(X\leqslant x_0)
\end{aligned} \tag{56}$$

However since $c+\varepsilon>c$, it follows from theorem B.2 that $E[X^{c+\varepsilon}]=\infty$. Now applying lemma B.1 with $a=x_0$, we obtain that $E[X^{c+\varepsilon}|X>x_0]=\infty$. This means that the right hand side of equation 56 is infinite since the term $E[e^{sX}|X\leqslant x_0]\,P(X\leqslant x_0)$ is bounded below by 0. Therefore, for all $s>0$, $M_X(s)=\infty$ from which the claim follows immediately. □

Corollary B.5. *There is no open interval around 0 where the moment generating function of a heavy-tailed random variable converges.*

Proof. By defintion B.2, a heavy-tailed random variable has tail coefficient strictly less than 2. Thus by theorem B.4, the claim follows. □

Definition B.3. *A random variable X is called Pareto with minimum value x_m and scale parameter α if there exist two positives real numbers x_m and α such that:*

$$F_X(x) = \begin{cases} 0, x < x_m \\ 1 - \left(\frac{x_m}{x}\right)^{\alpha}, x \geqslant \alpha \end{cases}$$

where $F_X(x) = P(X \leqslant x)$ is the CDF of X.

Theorem B.6. *A Pareto random variable with scale parameter α has tail coefficient $c = \alpha$.*

Proof.

$$E[X^n] = \int_{x=-\infty}^{\infty} x^n dF_X = \int_{x=-\infty}^{\infty} x^n \frac{dF_X}{dx} dx = \int_{x_m}^{\infty} x^n \frac{\alpha x_m^{\alpha}}{x^{\alpha+1}} dx = \alpha x_m^{\alpha} \int_{x_m}^{\infty} x^{n-\alpha-1} dx$$

Thus, $E[X^n] < \infty$ if and only if:

$$n - \alpha - 1 < -1 \Leftrightarrow n < \alpha$$

As a result, the tail coefficient c of X is equal to α. □

Corollary B.7. *A Pareto random variable with scale parameter α is heavy-tailed (in the sense of definition B.2) if and only if $\alpha < 2$.*

Proof. Immediate consequence of theorem B.6. □

C Optimal twisting of the M/M/1 queue

In section 3.2, we have established an asymptotically optimal change of measure for certain classes of G/G/1 queues called exponential twisting. In this appendix, we examine the application of exponential twisting to the M/M/1 queue. This is motivated in part because the analytical simplicity of the M/M/1 model makes for an elegant exposition, but also and primarily because of the popularity of the M/M/1 model. Not that the two characteristics are unrelated.

By definition, an M/M/1 queue is a single-server queue with exponentially distributed service time and exponentially distributed inter-arrival time. If the mean service time is μ^{-1} and the mean inter-arrival time is λ^{-1}, then by definition 3.1, the inter-arrival times and service times are admissible if:

$$\mu > \lambda$$

Reusing the nomenclature from chapter 3, the inter-arrival probability distribution is f_A is given by:

$$f_A(x) = \lambda e^{-\lambda x} \tag{57}$$

and the service time distribution, f_B is given by:

$$f_B(x) = \mu e^{-\mu x} \tag{58}$$

It can easily be checked from equations 57 and 58 that the moment generating function of $X = B_1 - A_0$ is:

$$M_X(\theta) = M_B(\theta) M_A(-\theta) = \frac{\mu}{\mu - \theta} \cdot \frac{\lambda}{\lambda + \theta} \tag{59}$$

where $M_B(\theta)$ and $M_A(\theta)$ are the moment generating functions of the service times and inter-arrival times respectively. Solving the equation $M_X(\theta^*) = 1$ for positive roots then yields:

$$\theta^* = \mu - \lambda$$

Now let f_A^θ be the θ-twisted distribution of the inter-arrival times:

$$f_A^\theta(x) = \frac{f_A(x) e^{\theta x}}{M_A(\theta)}$$

and similarly, let f_B^θ be the θ-twisted distribution of the service times:

$$f_B^\theta(x) = \frac{f_B(x)\, e^{\theta x}}{M_B(\theta)}$$

It is then straightforward to check that:

$$f_A^{\theta^*}(x) = f_B(x)$$

and

$$f_B^{\theta^*}(x) = f_A(x)$$

In other words, the asymptotically optimal exponential change of measure from section 3.2 corresponds to simulating another M/M/1 queue with the service and arrival rates exchanged. Although in light of corollary 3.7 (page 39), it should not be surprising that the exponentially-twisted queue is unstable, it may not intuitively be obvious why exchanging service and arrival times in an M/M/1 queue is the optimal importance sampling change of measure (within the class of random walks with i.i.d. increments).

Some intuition for this result is provided in [1] where it is shown that, conditioned on an M/M/1 queue having taken a large value, it is almost surely the case that the queue was stable up to a certain time, and then started growing at a rate $\mu - \lambda$ until it reached the large value. For a more formal exposition, see [1], but this illustrates that a good importance sampling measure, and indeed an asymptotically optimal one in this case, is one that concentrates the probability mass of the sampling distribution over the simulation path that the rare event (in our case, the queue exceeding some specified size n) is most likely to have taken.

References

[1] V. Anatharam. How large delays build up in a gi/g/1 queue. *Queueing Systems*, 5:345–368, 1988.

[2] S. Asmussen and K. Binswanger. Simulation of ruin probabilities for subexponential claims. *Astin Bulletin*, 27(2):297–318, 1997.

[3] Soren Asmussen. *Applied Probability and Queues*, chapter 8, page 237. Springer, 2nd edition, 2003.

[4] Dimitri Bertsekas and Robert Gallager. *Data Networks*, chapter 3.3, page 128. Prentice-Hall, 1987.

[5] Patrick Billingsley. *Probability and Measure*, chapter 34. Wiley-Interscience, 3rd edition, 1995.

[6] Luc Devroye. *Non-Uniform Random Variate Generation*, chapter 2, pages 27–28. Springer-Verlag, 1986.

[7] Geofrrey Grimmet and David Stirzaker. *Probability and Random Processes*, chapter 11.5, page 455. Oxford University Press, 3rd edition, 2001.

[8] Geofrrey Grimmet and David Stirzaker. *Probability and Random Processes*, chapter 5.1, page 150. Oxford University Press, 3rd edition, 2001.

[9] Allan Gut. *Stopped Random Walks: Limit Theorems and Applications*, chapter 3.1, page 80. Springer, 2nd edition, 2009.

[10] Daniel P. Heyman and Teunis J. Ott. On the choice of alternative measures in importance sampling with markov chains. *Operations Research*, 43(3):509–519, May-June 1995.

[11] Sandeep Juneja and Perwez Shahabuddin. Simulating heavy tailed processes using delayed hazard rate twisting. *ACM Transactions on Modelling and Computer Simulation*, 12(2):94–118, April 2002.

[12] T. Lehtonen and H. Nyrhinen. Simulating level-crossing probabilities by importance sampling. *Advanced Applied Probability*, 24:858–874, 1992.

[13] Natalia Markovich. *Nonparametric analysis of univariate heavy-tailed data: research and practice*, chapter 1.2. Wiley-Interscience, 1st edition, 2007.

[14] Gabriel Nagy. *Real Analysis*, chapter 4.4, pages 300–321. http://www.math.ksu.edu/~nagy/real-an/.

[15] S. I. Resnick. Heavy tail modeling and teletraffic data. *The Annals of Statistics*, pages 1805–1869, 1997.

[16] D. Siegmund. Importance sampling in the monte carlo study of sequential tests. *Annals of Statistics*, 4:673–684, 1976.